Novel Uses of Epigenetics in Forensic Science

Vidaki, Athina

Awarding institution:
King's College London

Novel Uses of Epigenetics in Forensic Science

Athina Vidaki MSc

January 2015

A thesis submitted to the University of London

for the degree of Doctor of Philosophy

King's College London

University of London

Abstract

Body fluids such as blood are amongst the most important biological evidence recovered from crime scenes. Identification of the donor can be achieved through STR profiling; however, extracting additional information regarding the tissue type or the donor's physical appearance such as age could prove very useful in police investigations. Firstly, the performance of existing tissue-specific mRNA-based systems was assessed via collaborative exercises. All proposed methods have shown to be highly sensitive; however, issues regarding markers' specificity, especially for the vaginal detection, were observed. Analysing complex casework samples revealed the need for interpretation guidelines and the use of a scoring system when implementing mRNA profiling in casework. It was understood that developing DNA-based testing would overcome the limitations of existing methods so the main aim of this study was to evaluate the applicability of DNA methylation profiling in forensics.

Using three approaches various tissue-specific differentially methylated CpG sites in 18 different loci were evaluated by analysing various forensically relevant body fluids and tissues. As a result, a set of suitable blood- and semen-specific markers were validated using aged and mock casework samples; however, the identification of other tissues like saliva, vaginal fluid and menstrual blood seemed to be challenging. Regarding age prediction, a set of age-associated CpG sites were selected from genome-wide DNA methylation studies and the correlation of their blood methylation levels with age was assessed on two sequencing platforms. Using a subset of 16 CpG sites and taking advantage of artificial neural networks' capabilities, age could be accurately predicted in 1,156 blood samples (mean error of 4.1 years). The applicability of the proposed prediction model was also tested by means of next generation sequencing. Although further research is required prior to implementing these results in casework, it can be concluded that epigenetics could shed light on the proposed forensic applications.

Acknowledgments

I would like to thank everyone that has been involved in this research throughout the last four years without whom I would not be able to get to this stage.

First of all, I would like to thank the Papadaki Foundation and National and Kapodistrian University of Athens for funding my studies since 2009, without which conducting this study would have never been possible. I would also like to thank King's College London for their financial support in times where I needed it the most.

A huge thank you to Denise Syndercombe Court who has been the best and most supportive supervisor I could have ever wished for. Thank you for believing in me and for always keeping your door open. I will always be grateful for the opportunities you gave me along the last four years towards my development as an independent researcher and forensic scientist.

I would also like to thank Barbara Daniel for initiating the project and her advice throughout as well as Leon Barron for his valuable help with data analysis. Special thanks to David Ballard for his advice on experimental design without which I would often find myself lost! I am extremely grateful for your help and contribution and for always being there for me.

I would like to thank everyone that have advised me both at Queen Mary and King's College who are probably too many to list here but I am very much thankful for their contribution. A big thank you to all my volunteers as well as Lawrence for assisting in sample collection. I would also like to thank everyone involved in the European collaborative projects that have made it a great experience.

A big high-five to my friends and colleagues in the lab, especially Chun-Wai, Gabriella, Cecilia, Federica, James, Nacho, Andrew and Lesley who made sure PhD is the best experience in my life so far. Special thanks to Nanda, I hope you know how lucky I am to have you in my life and how appreciative I am for your advice as well as the numerous distractions that helped me come out sane!

Finally, I would like to thank a million my mother and sister, for their continuous support and faith and for always believing I can succeed. I know you have both been through hard times to make sure I accomplish my dreams and I will always be grateful. Lastly, I would like to dedicate this work to my beloved father who is not here to witness my progress, of which I am sure he would be very proud.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A	Adenine
AAR	Aspartic Acid Racemisation
ACTB	Actin Beta
ACVR1	Activin A Receptor Type 1
AGE	Advanced Glycation End-Product
AIDS	Acquired Immune Deficiency Syndrome
ALAS2	Δ-Aminolevulinate Synthase 2
AMICA1	Adhesion Molecule, Interacts With CXADR Antigen 1
ANK1	Ankyrin
ANKRD34C	Ankyrin Repeat Domain 34C
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
AP	Acid Phosphatase
APS	Adenosine 5' Phosphosulfate
AQP9	Aquaporin 9
ASPA	Aspartoacylase
ATP	Adenosine Triphosphate
ATP13A2	Transmembrane Lysosome P5-Type ATPase
Avag	*Atopobium Vaginae*
B2M	Beta-2 Microglobulin
BCAS4	Breast Carcinoma Amplified Sequence 4
BIRC4BP	BIRC4-Binding Protein
BLAST	Basic Local Alignment Search Tool
BLM1	Blood Methylated Marker 1
BLMH	Bleomycin Hydrolase
BLU1	Blood Unmethylated Marker 1
BLU2	Blood Unmethylated Marker 2
bp	Base Pairs
BS	Bisulphite Sequencing
BUM1	Buccal Cell Methylated Marker 1
BUM2	Buccal Cell Methylated Marker 2
BUU1	Buccal Cell Unmethylated Marker 1
BUU2	Buccal Cell Unmethylated Marker 2
BWA	Burrows-Wheeler Alignment
C	Cytosine
C19orf30	Hypothetical Protein LOC284424
C20orf117	Chromosome 20 Open Reading Frame 117
C21orf63	Chromosome 21 Open Reading Frame 63
cAMP	Cyclic Adenosine Monophosphate
CARD15	Caspase Recruitment Domain-Containing Protein 15
CAS	CRISPR-Associated
CASC4	Cancer Susceptibility Candidate 4 Isoform A
CCD	Charge Coupled Device
Ccer1	Coiled-Coil Glutamate-Rich Protein 1

CCL27	Chemokine Ligand 27
CCO	Cytochrome C Oxidase
CD248	Cluster of Differentiation 248
CD3	Cluster of Differentiation 3
CD3G	CD3g Molecule
CD4	Cluster of Differentiation 4
CD8b	Cluster of Differentiation 8b
CD93	Cluster of Differentiation 93
CDKN2A	Cyclin-Dependent Kinase Inhibitor 2A
CDM	Cell Type Specific Differentially Methylated Gene Region
cDNA	Complementary DNA
CDSN	Corneodesmosin
CE	Capillary Electrophoresis
C-glyTrp	C-Glycosyl Tryptophan
Ck1e	Casein Kinase 1
CNV	Copy-Number Variation
CORBA	Combined Bisulphite Restriction Analysis
CpG	Cytosine-Phosphate-Guanine
Cry1	Cryptochrome Circadian Clock 1
Cry2	Cryptochrome Circadian Clock 2
CSNK1D	Casein Kinase 1, Delta Isoform 1
CSNK1D	Casein Kinase 1
CYP2B7P1	Cytochrome P450, Family 2, Subfamily B, Polypeptide 7 Pseudogene 1
DACT1	Dapper 1 Isoform 2
DAPK1	Death-Associated Protein Kinase 1
DMAC	Dimethylaminocinnamaldehyde
DMR	Differentially Methylated Region
DNA	Deoxyribonucleic Acid
DNMT1	DNA Methyltransferase 1
DNMT3b	DNA Methyltransferase 3b
dNTP	Deoxyribonucleotide Triphosphate
DPD	Deoxypyridinoline
dsDNA	Double-Stranded DNA
DTT	Dithiothreitol
EDARADD	Edar-Associated Death Domain
EDNAP	European DNA Profiling Group
EDTA	Ethylenediaminetetraacetic Acid
EFS	Embryonal Fyn-Associated Substrate
ELK	ETS-Domain Protein
ELOVL2	ELOVL Fatty Acid Elongase 2
ERG	V-Erythroblastosis Virus E26 Oncogene Like Isoform 2
ESR1	Estrogen Receptor 1
EuroForGen	European Forensic Genetics Network Of Excellence
FACT	Facilitates Chromatin Transcriptional Elongation Factor
FDCSP	Follicular Dendritic Cell Secreted Protein
FGF7	Fibroblast Growth Factor 7
FHL2	Four And A Half LIM Domains 2
FIA-MS/MS	Flow Injection Analysis/Mass Spectrometry

FXN	Frataxin, Mitochondrial Isoform 1 Preprotein
FZD9	Frizzled 9
G	Guanine
G6PDH	Glucose-6-Phosphate Dehydrogenase
GAPDH	Glyceraldehyde-3-Phosphatedehydrogenase
GATK	Genome Analysis Tool kit
GlycoA	Glycophorin A
GOLM1	Golgi Membrane Protein 1
GRIA2	Glutamate Receptor, Ionotropic, AMPA 2
GRNN	Generalised Regression Neural Network
Gvag	*Gardnerella vaginalis*
HBA	Haemoglobin Alpha
HBA1	Haemoglobin Alpha 1
HBB	Haemoglobin Beta
HBD1	Human Beta-Defensin 1
HIV	Human Immunodeficiency Virus
HKG	Housekeeping Gene
HOXA4	Homeobox Protein Hox-A4
HRM	High Resolution Melting Analysis
HS	High Sensitivity
HTA	Human Tissue Act
HTERT	Human Telomerase Reverse Transcriptase
HTN1	Histatin Isoform
HTN3	Histatin 3
HymA	Hydrogenase Subunit Hyma
IL19	Interleukin 19
IL1F7, IL37	Interleukin 1 Family Member 7
IL6	Interleukin 6
IPC	Internal Positive Control
ITGA2B	Integrin Alpha 2b
KCNQ1DN	Potassium Voltage-Gated Channel, KQT-Like Subfamily, Member 1 Downstream Neighbour (Non-Protein Coding)
KLF14	Kruppel-Like Factor 14
KLK3	Kallikrein 3
KM	Kastle-Mayer
KRT13	Keratin 13
KRT4	Keratin 4
KRT6A	Keratin 6A
KRT9	Keratin 9
LCE	Late Cornified Envelope Gene
LCE1C	Late Cornified Envelope Gene 1C
LCE1D	Late Cornified Envelope Gene 1D
LCE2D	Late Cornified Envelope Gene 2D
LCL	Lymphoblastoid Cell Line
LCN2	Lipocalin 2
LEFTY2	Left-Right Determination Factor 2
LGM	Leucomalachite Green
Liners	*Lactobacillus iners*

Ljen	*Lactobacillus jensenii*
LOR	Loricrin
LZTR1	Leucine-Zipper-Like Transcription Regulator 1
MAD	Mean Absolute Deviation
mC	Methylated Cytosine
MGB	Minor Groove Binder
MIR649	Mirna Gene 649
miRNA	Micro RNA
MLP	Multi-Layer Perceptron
MMP 10	Matrix Metalloproteinase 10
MMP 11	Matrix Metalloproteinase 11
MMP 7	Matrix Metalloproteinase 7
MR	Methylation Ratio
mRNA	Messenger RNA
MSI2	Musashi RNA-Binding Protein 2
MSLN	Mesothelin
MSNB	Beta-Microseminoprotein
MS-PCR	Methylation-Specific PCR
MS-RDA	Methylation-Sensitive Representational Difference Analysis
MSRE-PCR	Methylation-Specific Restriction Enzyme PCR
MS-SNuPE	Methylation-Sensitive Single-Nucleotide Primer Extension
MSX	Mshhomeobox
MSX1	Mshhomeobox 1
mtDNA	Mitochondrial DNA
MUC4	Mucin 4
MUC7	Mucin 7
MYOZ1	Myozenin 1
NCBI	National Center For Biotechnology Information
NDNAD	National DNA Database
NFI	Netherlands Forensic Institute
NFKB1	Nuclear Factor Kappa-B Subunit 1
NFQ	Non-Fluorescent Quencher
NGS	Next Generation Sequencing
NHLRC1	Malin
NK	Natural Killer Cell
NPTX2	Neuronal Pentraxin II
NR2F2	Nuclear Receptor Subfamily 2, Group F, Member 2
OG	Obligatory Gene
OSBPL5	Oxysterol Binding Protein-Like 5
P2RX6	P2X Purinoceptor 6
P2RXL1	Purinergic Receptor P2X-Like 1
PacBio	Pacific Biosciences
PARP6	Poly (ADP-Ribose) Polymerase Family, Member 6
PBGD	Porphobilinogen Deaminase
PBMC	Peripheral Blood Mononuclear Cell
PCR	Polymerase Chain Reaction
PDE4C	Phosphodiesterase 4C, Camp Specific
PDGFRA	Platelet-Derived Growth Factor Receptor Alpha

PEG3	Paternally Expressed 3
PENK	Proenkephalin
Per1	Period Circadian Protein Homolog 1
Per2	Period Circadian Protein Homolog 2
Per3	Period Circadian Protein Homolog 3
PFN3	Profilin-3
PGK1	Phosphoglycerate Kinase 1
PGM	Personal Genome Machine
PIA	Parentally Imprinted Allele
PKM2	Pyruvate Kinase 2
PMI	Post-Mortem Interval
PPBP	Pro-Platelet Basic Protein
PPi	Pyrophosphate
PPIA	Cyclophilin A
PRB1 - 4	Proline-Rich Proteins 1 - 4
PRM1	Protamine 1
PRM2	Protamine 2
PRMT2	Protein Arginine N-Methyltransferase 2
PSA	Prostate-Specific Antigen
qPCR	Quantitative PCR
RASSF5	RAS Association Domain Family 5 Isoform B
RBFN	Radial Basis Function Network
RIN1	Ras and Rab Interactor 1
RMSE	Root-Mean-Square Deviation
RNA	Ribonucleic Acid
RNase	Ribonuclease
RNA-Seq	RNA Sequencing
RPLP0	Ribosomal Protein P0
RPS15	Ribosomal Protein S15
RPS18	Ribosomal Protein S18
rRNA	Ribosomal RNA
RT	Reverse Transcription
RT-PCR	Reverse Transcription PCR
SCGN	Secretagogin Precursor
SD	Standard Deviation
SDHB	Succinate Dehydrogenase (Ubiquinine) Iron-Sulfur Subunit, Mitochondrial
SEM1	Semen Methylated Marker 1
SEM2	Semen Methylated Marker 2
SEMG1	Semenogelin 1
SEMG2	Semenogelin 2
SERP4	Secreted Frizzled-Related Protein 4
SEU1	Semen Unmethylated Marker 1
SEU2	Semen Unmethylated Marker 2
SH3	Src Homology 3
sjTREC	Signal Joint TCR Excision Circles
SLC25A31	Solute Carrier Family 25 (Mitochondrial Carrier; Adenine Nucleotide Translocator), Member 31
SLC6A4	Solute Carrier Family 6 (Neurotransmitter Transporter)

SLC7A4	Solute Carrier Family 7, Member 4
SNORD63.3	Small Nucleolar RNA SNORD63.3
SNP	Single Nucleotide Polymorphism
SNRPN	Small Nuclear Ribonucleoprotein Polypeptide N
SPM	Spermine
SPRR1A	Small Proline-Rich Protein 1A
SPRR3	Small Proline-Rich Protein 3
SPTB	B-Spectrin
SSRP1	Structure Specific Recognition Protein 1
SST	Somatostatin
STATH	Statherin
STR	Short Tandem Repeat
T	Thymine
TCR	T-Cell Receptor
tDMR	Tissue-Specific Differentially Methylated Region
TEF	Translation Elongation Factor-1a
TGFβRII	Transforming Growth Factor-B
TGM4	Transglutaminase 4
THAP7	Thanatos-Associated Protein 7
Tim	Timeless
TMEM151A	Transmembrane Protein 151A
TOM1L1	Target Of Myb1 (chicken)-Like 1
TRIM58	Tripartite Motif Containing 58
TRIP10	Thyroid Hormone Receptor Interactor 10
tRNA	Transfer RNA
UCE	Ubiquitin-Conjugating Enzyme E2D 2
UKAS	United Kingdom Accreditation Service
USP49	Ubiquitin Carboxyl-Terminal Hydrolase 49
UV	Ultraviolet
VGF	Nerve Growth Factor Inducible Precursor
VNTR	Variable Number Tandem Repeat
WGBS	Whole-Genome Bisulphite Sequencing
ZC3H12D	Zinc Finger CCCH-Type Containing 12D

1 Introduction

Body fluids such as blood and saliva are amongst the most important biological evidence found at crime scenes. In the field of forensic biology, scientists analyse body fluid stains in order to identify the donor and potentially incriminate or exclude suspects. It is known that DNA profiling is one of the most powerful tools of forensic scientists. Although 99.9% of human DNA sequences are the same amongst individuals, there are regions that vary and these can be used to distinguish one person from another (unless they are monozygotic twins). Scientific developments over the last few years have extended the amount of information that can be extracted even from minute amounts of evidence. These days hundreds of different loci can be successfully amplified using picogram or nanogram of starting DNA material and this information can be applied in criminal investigations. DNA profiling is employed for almost every type of body fluid stain or swab that is submitted as part of a criminal investigation. In cases where there is a suspect, the DNA profile obtained from a crime scene sample is compared against the suspect's DNA profile. However, in cases where there is no match or there is no suspect, the obtained DNA profile is uploaded to the DNA database in order to search for a potential match with previously stored profiles.

According to the Home Office's 2013 annual report, there is a 61% chance that the database will produce a possible match every time a DNA profile derived from a crime scene is searched against the National DNA Database (NDNAD), and of these, 40% are integral to solving the case (National DNA Database Strategy Board, 2012-2013). This is one of the highest detection rates in Europe, however most of these crimes refer to burglaries and vehicle crimes and there are still a large number of homicides and rape cases that remain undetected, unsolved, or there is simply not enough evidence for conviction. In cases where a DNA profile is obtained from a crime scene but there is no match to profiles on the NDNAD, police officers reach a dead end in their investigation.

DNA profiles notwithstanding the examination of recovered biological material can shed light on the identity of the donor and help investigators develop a more detailed and clear picture of the crime. It would be very advantageous if forensic scientists were able to support DNA findings either by linking a DNA profile with its tissue source or by extracting additional information that could assist criminal investigations. Firstly, the presence of a specific body fluid could potentially indicate a particular crime; for example, the detection of semen could indicate sexual assault. Current testing does

allow for the identification of the cellular tissue of origin; however, these methods are mostly protein-based and are used in a presumptive manner (Virkler & Lednev, 2009). They are presumptive because they lack specificity and sensitivity and there are some body fluids such as menstrual blood that have no such tests. Secondly, researchers have already determined genetic variations associated with ethnicity or hair, skin and eye colour and since these markers are of an individual's physical characteristics they could prove useful in police investigations (Bouakaze *et al.*, 2009; Spichenok *et al.*, 2011; Walsh *et al.*, 2013). However, there are traits that have not been studied to any great depth such as chronological age and such information has been highlighted as important in criminal investigations.

The aim of this research is to investigate genetic and epigenetic differences in forensically relevant body fluids and to devise accurate methods to not only identify the tissue source of a stain but also estimate the chronological age of the donor. Towards achieving these goals, it is believed that methods based on the analysis of nucleic acids surpass protein tests as they are more sensitive and have a higher discriminatory power. Furthermore, since DNA is more robust than proteins, this approach will be applicable to aged and degraded material. Current forensic DNA tests used for applications such as identification of an individual or for ethnicity prediction are based on differences in DNA sequence, such as short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs) (Bouakaze *et al.*, 2009; Butler, 2012). Since all cells throughout an individual's lifetime contain the same DNA (apart from random mutations), methods based on these variations would not be suitable in this project.

In this study, emphasis will be given to one of the known chemical modifications of DNA that occurs naturally, known as DNA methylation; others include histone acetylation and chromatin modelling. These modifications have been implicated in the regulation of transcription and translation and jointly come under the umbrella term of 'epigenetics'. Thus, this chapter will present an overview of epigenetics with an emphasis on DNA methylation and also shed light on existing forensic applications of DNA methylation profiling. Later, the state of recent research in the area of tissue identification and age estimation will be reviewed in Chapters 3 and 6 respectively.

1.1 Epigenetics

The character of each cell is determined by its protein components, which are the product of specific gene expression patterns; these patterns are determined during cell differentiation and carefully controlled. Critical determinants are DNA-binding transcription factors that manage a gene's activation or repression by interacting with specific DNA sequences in its promoter region (Svingen & Tonissen, 2006). This interaction between transcription factors and DNA generates a sequence of events which involves modification of chromatin structure leading to the construction of an active transcription complex (Cosma *et al.*, 1999). However, transcription factors are not themselves sufficient to control gene expression, other mechanisms are required to make those recognition sequences available or not for binding. It is known that DNA is not 'naked' but exhibits various chemical modifications; these chemical groups together with the way DNA is packaged are the ones that control the access to regulatory molecules.

1.1.1 The epigenome

Epigenetics refers to the study of heritable alterations in gene function or cellular phenotype caused by mechanisms other than changes in the DNA sequence itself [Figure 1-1]. It involves functionally relevant modifications, such as DNA methylation and histone modifications, both of which play a significant role in regulating gene expression without altering the DNA sequence. The molecular basis of epigenetics is complex and primarily involves alterations in the activation of particular genes. Also, the chromatin proteins associated with DNA may be activated or silenced and therefore ensure that each cell expresses only the genes that are necessary for an activity (Bird, 2007; Reik, 2007). The term 'epigenome' is a parallel to the word 'genome' and refers to the overall epigenetic status of a cell.

Epigenetic patterns are preserved during cell division just as the DNA sequence is inherited from one generation to the next; however, they change over an individual's lifetime (Bird, 2002). Epigenetic changes have been observed to occur in response to environmental exposure and can be affected by various factors such as diet and smoking (Rando & Verstrepen, 2007). Imprinting, gene silencing, X chromosome inactivation, reprogramming and carcinogenesis are all examples of epigenetic

processes. A very important cell function that is regulated by epigenetic mechanisms in mammals is cell differentiation, where stem cells become fully differentiated cells during embryogenesis (Rando & Verstrepen, 2007).

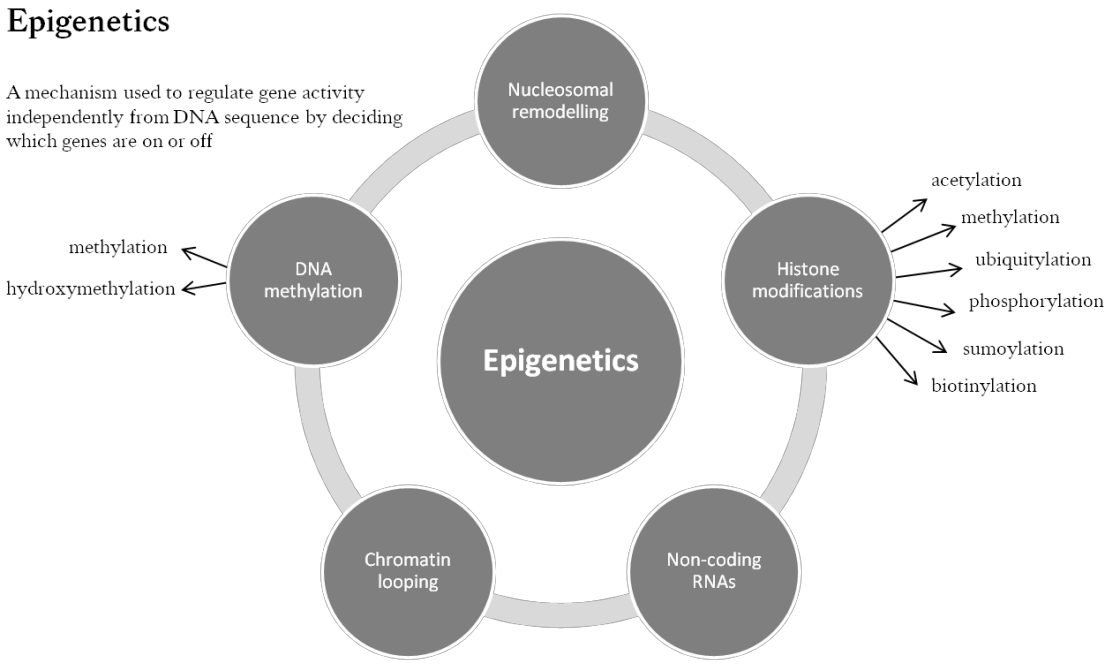


Figure 1-1. Epigenetic mechanisms (Vidaki *et al.*, 2013)
The field of epigenetics involves various mechanisms including DNA methylation and histone modifications; each mechanism can further involve different chemical alterations, such as methylation and hydroxymethylation indicated with arrows.

1.1.2 DNA methylation and gene regulation

In the human genome, DNA methylation is a vital biochemical process for normal development. It involves the addition of a methyl group (-CH_3) at the 5' position of cytosine residues in CpG dinucleotides [Figure 1-2a]. The absence of CpG methylation has only been observed in embryonic stem cells (Dogde *et al.*, 2002). Considering the genome as a whole it can be seen that most CpGs (60-90%) are methylated, whereas the unmethylated CpGs are often found grouped in areas known as "CpG islands" (300-3000 bp long, >55% GC content) usually located around the regulatory region (5' end) of many human genes (Espada & Esteller, 2010).

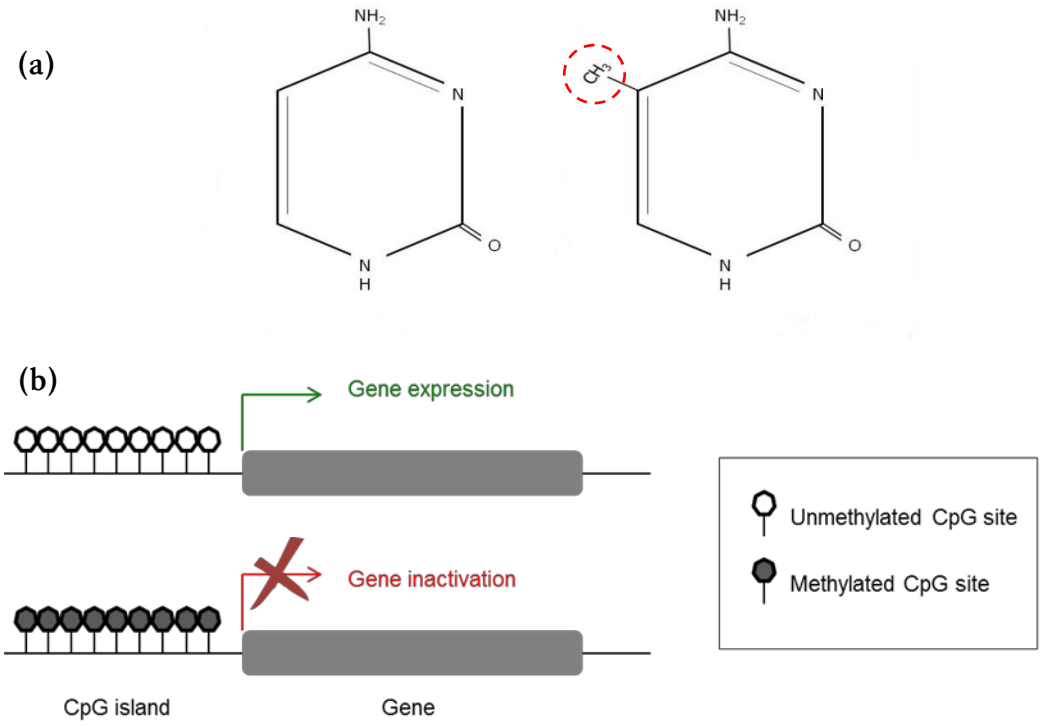


Figure 1-2. DNA methylation and gene regulation (Vidaki *et al.*, 2013)
(a) DNA methylation on cytosine, (b) schematic representation of a typical gene containing a CpG island; in most cases, when a CpG island is unmethylated allows for gene expression, while when it is methylated the gene becomes inactive.

DNA methylation occurs via DNA methyltransferases, which have two enzymatic activities: maintenance of methylation after every cellular DNA replication cycle and *de novo* methylation to set up new DNA methylation patterns early in development (Fernandez *et al.*, 2012). It is generally believed that DNA methylation found in the promoter regions of genes is associated with condensed nuclease-resistant heterochromatin and silencing of gene expression (Newell-Price *et al.*, 2000) [Figure 1-2b]. Methylation of DNA is followed by subsequent recruitment of binding proteins that preferentially recognize methylated DNA. As a result, histone deacetylase and chromatin remodelling complexes further stabilize the condensed chromatin. Conversely, the opposite is also possible (Newell-Price *et al.*, 2000). There have been cases in which DNA methylation in gene promoter areas has been associated with gene activation (Straussman *et al.*, 2009). Straussman and co-workers identified at least 50 loci where host genes are expressed exclusively in the same cell-type only when their promoters or gene bodies are methylated. This is mainly observed in non-CpG-island DNA and it is believed to be a result of a process of selective demethylation of inactive genes (Straussman *et al.*, 2009).

1.1.3 DNA methylation and the environment

Recent studies support the hypothesis that DNA methylation acts as an interphace between the fixed genome and the dynamic environment. Changes in DNA methylation patterns in response to environmental stress in early life or subsequently serve as a mechanism of life-long genome adaptation, enabling the same genome to express different phenotypes (Szyf, 2010). Several studies examining monozygotic twins have established a link between environment or ageing and long-lasting epigenetic impressions on phenotype (Fraga *et al.*, 2005; Wong *et al.*, 2005). Monozygotic twins serve as the ideal system to investigate epigenetics as they share the same genetic basis. It has been reported that twins at a young age show similar amounts of DNA methylation, whereas as they grow older they differ significantly not only in the amount but also in the patterns of DNA methylation (Fraga *et al.*, 2005) [Figure 1-3]. Scientists have hypothesized that these non-genetic age-dependent differences in 'gene marking' could explain differences in a broad range of physical characteristics or disease susceptibility, often observed in monozygotic twins (Fraga *et al.*, 2005).

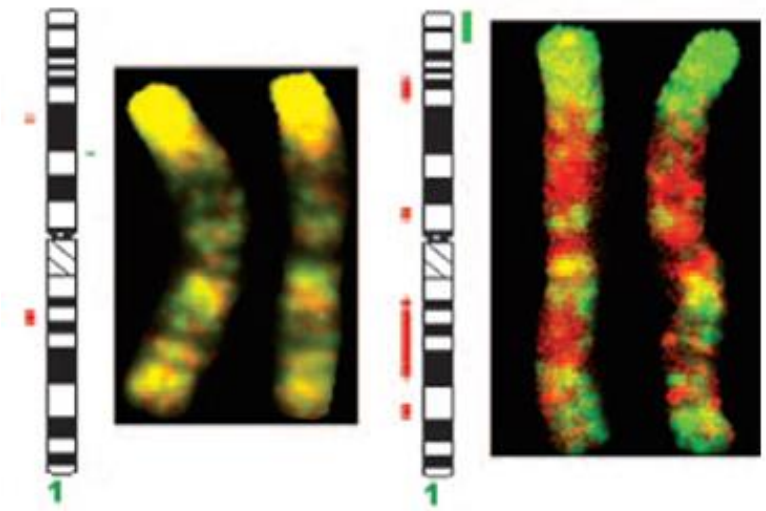


Figure 1-3. Mapping chromosomal regions with differential DNA methylation in two pairs of monozygotic twins of different ages (Fraga *et al.*, 2005)
Competitive genomic hybridization for methylated DNA onto normal metaphase chromosomes 1 was generated from 3- and 50-year-old twin pairs. Green signals indicate hypermethylation, whereas red signals show hypomethylation events. On the other hand, the yellow colour is obtained by equal amounts of the green and red dyes. Therefore, the 50-year-old twin pair shows plentiful changes in the pattern of DNA methylation whereas the 3-year-old twins have a very similar distribution of DNA methylation.

1.1.4 DNA methylation as a biomarker

As mentioned above, it has been demonstrated that genes with high DNA methylation levels in their promoter regions are usually transcriptionally silent, and that DNA methylation gradually accumulates upon long-term gene silencing. Epigenetic mis-programming leading to aberrant DNA methylation patterns (hyper/hypo-methylation) is a significant observation in several human diseases and malignancies. Hypermethylation occurring at CpG islands in the promoter region of cancer-associated genes results in gene inactivation, however genome-wide hypomethylation has been associated with cancer progress and metastasis (Wild & Flanagan, 2010). Interestingly, there is evidence to support two competing hypotheses of how hypomethylation might occur; one includes a 'passive' loss of the maintenance of DNA methylation over cell divisions and the other involves an 'active', DNA replication-independent loss of methylation, which could potentially occur faster.

1.1.5 DNA methylation analysis

Determination of DNA methylation patterns either at single CpG sites, or the entire methylome has become increasingly important in medical diagnoses and different methods of detecting methylation have been developed. These are based on the a) chemical modification of the unmethylated cytosine residues, b) protein interaction with 5-methyl cytosine and c) digestion by methylation-sensitive restriction enzymes. In certain cases, depending on the scientific aim, a combination of these techniques may be employed (Ammerpohl *et al.*, 2009; Oakeley, 1999).

1.1.5.1 Chemical modification of cytosine residues

The methylation profile of a target DNA sequence can be established through treatment with sodium bisulphite ($NaHSO_3$) (Fraga & Esteller, 2002). During this process unmethylated cytosine residues are converted into uracil, while methylated cytosines remain unchanged. Thus, bisulphite treatment results in different DNA sequences for the methylated and unmethylated DNA [Figure 1-4]. Correct determination of a specific methylation pattern is highly dependent on the complete conversion of unmethylated cytosines. This procedure suffers, to a degree, from DNA fragmentation/loss (Tanaka & Okamoto, 2007), but is probably needed in order to obtain high conversion rates of the unmethylated cytosines.

Conversion of DNA using sodium bisulphite allows for both qualitative and quantitative analysis of individual CpG sites. However, the possibility of incomplete conversion of cytosines has to be taken into account as it could result in an overestimation of methylation. Following bisulphite treatment it is necessary to amplify the DNA using the polymerase chain reaction (PCR) prior to sequencing and assuming that all non-CpG cytosines have been converted into uracil, the DNA will mainly consist of three bases (DNA polymerase does not recognise uracil, therefore it is treated as thymine). Due to this reduced complexity, the options in primer design are limited and can result in low yields of PCR products or non-specific products due to mis-priming events. Also, differences in methylated and unmethylated DNA sequences may affect amplification efficiency by introducing PCR bias (Moskalev *et al.*, 2011; Warnecke *et al.*, 1997).

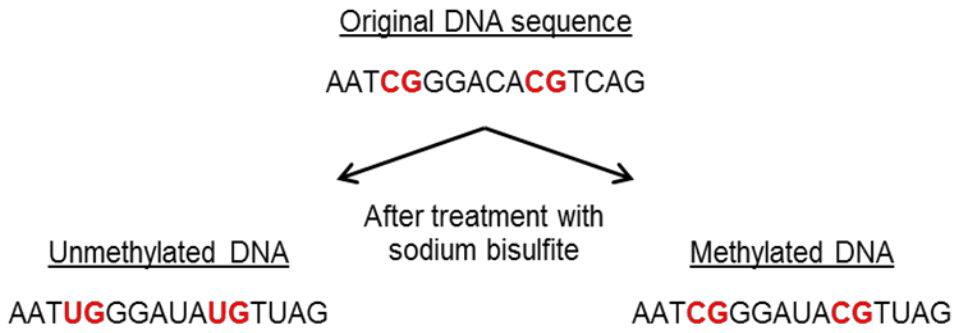


Figure 1-4. Bisulphite conversion of DNA
After treatment with sodium bisulphite only non-methylated cytosines are converted into uracil, while the methylated ones remain unchanged.

1.1.5.2 Protein interaction with 5-methyl cytosine

In contrast with bisulphite treatment, methods based on protein precipitation of DNA maintain the structure of DNA bases (Jacinto *et al.*, 2008); however, they lack high specificity and sensitivity. For instance, methylated DNA immunoprecipitation is a large-scale purification technique that is used to enrich for methylated DNA sequences. It mainly consists of isolating methylated DNA fragments via a monoclonal antibody raised against 5-methyl cytosine. To obtain the fragments, DNA is usually subjected to sonication and denaturation. The short length of these fragments is crucial in achieving sufficient resolution, improving the efficiency of downstream analysis as well as reducing fragment-length effects or bias. Such techniques can be combined with PCR or array based methods; nevertheless, caution must be taken as limited amounts of DNA could also result in PCR bias.

1.1.5.3 Methylation-sensitive restriction enzymes

Restriction enzymes whose cleavage activity depends on the methylation of a CpG site in their target sequence can also be used for methylation analysis (Hua *et al.*, 2011) [Figure 1-5]. These enzymes, also known as methylation sensitive endonucleases usually cut unmethylated DNA only. This approach is simple and robust but incomplete digestion or differences in enzyme activity can hinder the analysis and affect interpretation. A main drawback associated with the use of restriction enzymes is the dependence on the availability of specific recognition sequences that flank the CpG site of interest.

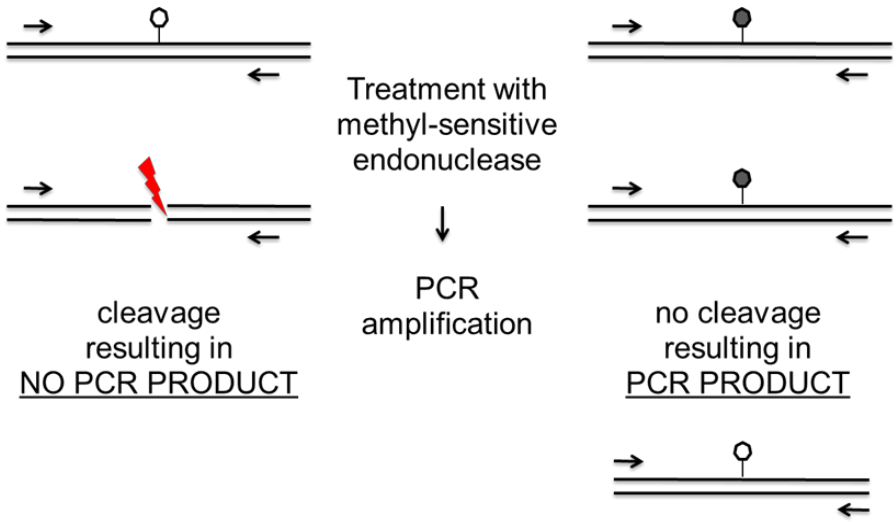


Figure 1-5. Methylation-sensitive restriction enzymes
Genomic DNA containing unmethylated (white circle) and methylated (black circle) CpG sites is restricted using a methylation-sensitive endonuclease. DNA is cleaved only when it is not 'protected' by a methyl group resulting in no PCR product after amplification with primers flanking the site in question (arrows). However, when DNA is methylated, the endonuclease cannot cut the recognition site leading to the generation of a PCR product. Note that the PCR product is shown as being unmethylated (white circle) since the incorporated nucleotides during PCR are always unmethylated; therefore, all methyl groups are lost after the first PCR cycle.

In the context of forensic analysis, it is believed that methods based on bisulphite treatment are the most appropriate as the integrity of DNA is maintained and the quantity of starting DNA material needed is relatively low. Alternative procedures may also be sensitive to external inhibitory compounds, commonly found in forensic casework. This is very important as forensic samples are frequently of low quality and quantity, aged or degraded. There have been various reported methods based on bisulphite treatment of DNA and will be discussed in Chapter 5.

1.2 Forensic Epigenetics

Various cellular decisions including survival, growth and differentiation are controlled by particular gene expression patterns, which are further regulated by changes in the epigenetic state of 'key' genes. The possibility of quantifying the DNA methylation levels of specific genes is of interest in a broad range of scientific and medical disciplines. As mentioned above, the epigenome acts as an interphace between the genome and the environment, and can change according to the specific needs of a cell. Therefore, DNA methylation patterns could reveal the activities of genes within a certain tissue at a certain point in time. From a forensic standpoint the analysis of DNA methylation patterns may give hints on pathological states (Das & Singal, 2004) or circumstances leading to death (Fragou *et al.*, 2011). Some of the main forensic applications of epigenetic analysis will be discussed below.

Although DNA methylation is a relatively new approach for gene expression studies, it has been previously reported in forensic research. Naito and his colleagues were the first to introduce DNA methylation in forensic genetics and proposed a method for epigenetic female sex typing (Naito *et al.*, 1993). The method was based on differential methylation pattern of an X chromosome-specific repetitive sequence, DXZ4 region, which was found highly methylated on the active X, whereas it was hypomethylated on the inactive X chromosome. This novel protocol for positively determining female sex was shown to be very sensitive as only 50 pg of DNA was required for successful sex typing, and the method had been proposed as a complementary technique in cases of sex-reversed individuals.

1.2.1 Determination of the parental origin of alleles

In paternity testing, the alleged father is usually excluded when he does not have the obligatory gene (OG), which can be determined from the child's genotype when the maternal gene is known. However, in motherless cases, the OG cannot be determined resulting in a reduced exclusion probability. Consequently, the determination of the parental origin of alleles without genealogical analysis could serve as a solution. We all inherit two copies (alleles) of each autosomal gene, one from our mother and one from our father. In most cases both alleles are active and functional; however, for a small group of genes one copy is turned off in a parent-of-origin dependent manner (Li *et al.*,

1993; Reik & Walter, 2001). The epigenetic mark of a gene based on its parental origin is called genomic imprinting and results in monoallelic expression. Maternal and paternal alleles are differentially methylated; therefore the parental origin of an allele could be determined by analysing its DNA methylation status. Imprinted expression can vary among tissues and developmental stages, while aberrant imprinting is the cause of various disease syndromes (Reik & Walter, 2001).

Zhao and his colleagues studied the methylation status at the imprinted SNP locus rs220028 and were able to determine the parental origin of alleles in all tissues tested (Zhao *et al.*, 2005). Their methodology involved a combination of digestion with methylation-sensitive restriction enzyme and post-digestion PCR using allele-specific primers thus allowing for the simultaneous analysis of the polymorphism and the methylation marker in one PCR reaction [Figure 1-7]. The authors advised that using too much DNA could lead to false methylation detection through incomplete digestion by the methylation-sensitive restriction enzyme. This technique is feasible for relationship testing but may prove a drawback for forensic samples.

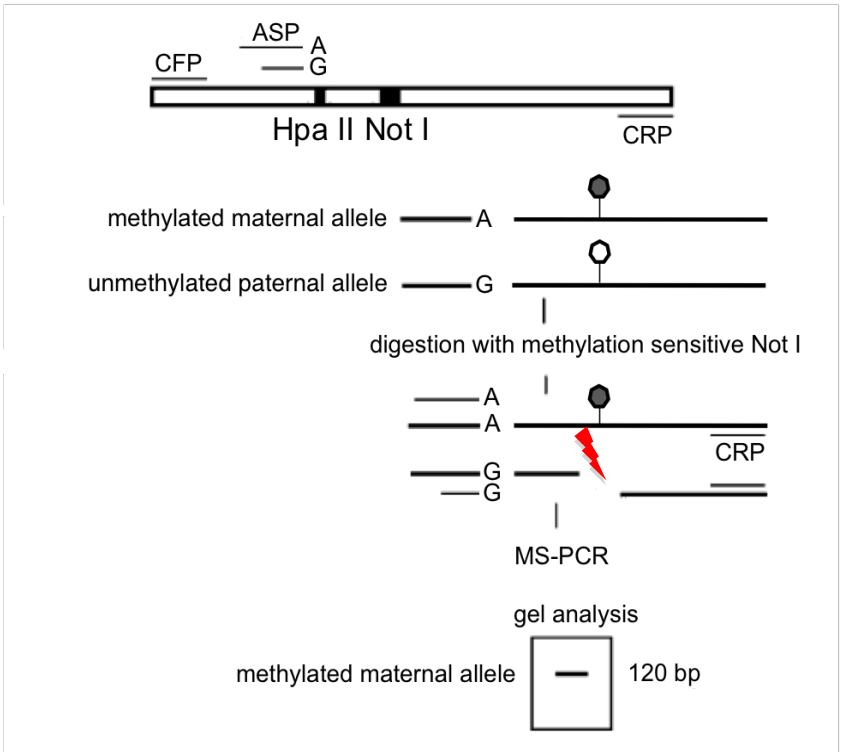


Figure 1-6. Strategy for determining the parental origin of an allele (Zhao *et al.*, 2005)
The structure of the imprinted SNP locus rs220028 is shown; the locus is found methylated (black circle) when inherited from the mother and unmethylated (white circle) when inherited from the father. Black boxes represent the recognition sites for enzymes Not I (5'- GC|GGCCGC -3') and Hpa II (5'- C|CGG -3') (also SNP site A/G). The primers used in the study are indicated as straight lines. The parental origin of an allele is determined by post-digestion methylation-specific PCR (MS-PCR).

Another study includes parentally-imprinted allele (PIA) typing of a variable number tandem repeat (VNTR) locus at the 5'-end of the *H19* imprinted gene, in blood samples and in a range of post-mortem tissues (Sumi *et al.*, 2005). Their protocol involved separate PCR for genotyping and PIA typing and showed no tissue-dependent methylation differences. The potential of *H19* gene in high-resolution paternity testing has also been investigated by other research groups (Huang *et al.*, 2008; Naito *et al.*, 2003) who reported two parentally imprinted H19FRs haplotype polymorphisms.

Furthermore, Nakayashiki and his team performed a more comprehensive study analysing a set of several SNPs in five imprinted genes using bisulphite sequencing in a range of tissues (Nakayashiki *et al.*, 2009a; Nakayashiki *et al.*, 2008; Nakayashiki *et al.*, 2009b). The maternal allele could be detected from the maternally-expressed gene (*H19*), and the paternal allele from the paternally-expressed ones: the hydrogenase subunit HymA (*HYMA1*), small nuclear ribonucleoprotein polypeptide N (*SNRPN*) and paternally expressed 3 (*PEG3*) gene (Nakayashiki *et al.*, 2008). However, unpredicted DNA methylation levels were detected in a few tissues, such as nails, hair and sperm as well as in samples exposed in higher temperatures or longer preservation periods (Nakayashiki *et al.*, 2009a; Nakayashiki *et al.*, 2009b).

In conclusion, the parental discrimination of a polymorphic allele in imprinted genes could be very promising for autosomal heritage analysis in difficult paternity cases although further research on tissue specificity and DNA methylation stability is required. Recent systematic, large-scale efforts to study these effects and to profile both tissue-shared and tissue-specific imprinting levels in humans have highlighted that there is indeed variation in imprinting not only between genes, but also between tissues and individuals (Baran *et al.*, 2015).

1.2.2 Authentication of DNA samples

DNA evidence has become an invaluable tool for forensic testing and personal identification. The use of DNA recovered from the crime scene and presented as evidence in court relies on the assumption that it is genuine, meaning originating from natural biological material. However, using well-established methods (Lasken & Egholm, 2003) DNA can be artificially produced and it is conceivable that DNA with

any desired genetic profile can be synthesized *in vitro* and planted in crime scenes (Frumkin *et al.*, 2010). Conventional genotyping techniques would fail to distinguish between natural and artificial DNA samples. Frumkin *et al* proposed a method of identifying artificial DNA based on differential DNA methylation patterns.

The authors synthesised DNA *in vitro* using three different methods which all produced unmethylated DNA. On the other hand, *in vivo* generated DNA contains both unmethylated and methylated loci. For the purpose of this study, two loci that are consistently methylated and two loci that are consistently unmethylated were chosen. Moreover, analysing some mock casework samples and natural/artificial DNA mixtures, it was shown that *in vitro* synthesized DNA could be detected, even in cases where it represented a minor component (Frumkin *et al.*, 2010).

Since synthesising DNA *in vitro* would require knowledge of molecular biology as well as dedicated instrumentation, planting someone else's biological material at a crime scene would definitely be an easier option. Furthermore, authors did not take into account that following its synthesis artificial DNA can be selectively methylated if desired by the use of specially engineered methyltransferases (Chaikind *et al.*, 2012). However, in cases where the presence of artificial DNA is suspected, the above method could be employed.

1.2.3 Discrimination of monozygotic twins

Individual identification of monozygotic twins in forensic casework has always been a challenge. Conventional forensic DNA typing cannot be performed, as monozygotic twins are expected to have the same DNA sequence. To further highlight how genetically identical twins are, a recent study performed ultra-deep next generation sequencing in a pair of twins to explore possible differences (Weber-Lehmann *et al.*, 2014). Researchers revealed only five extremely rare mutations in the whole genome; these results were also confirmed by classical Sanger sequencing. Although these results are interesting and quite promising, it is impractical to apply this approach in forensic casework due to the high cost and often, to the quality and quantity of the DNA available.

Twins do show various degrees of phenotypic difference (Fraga *et al.*, 2005; Wong *et al.*, 2005), which are presumably due to environmental differences acting on the

genome through epigenetic processes. Support for this theory comes from human disease studies, where evidence has been gathered that (apart from other mechanisms such as copy-number variation (CNV) in somatic cells) (Bruder *et al.*, 2008), epigenetic modifications are the ones that are responsible for these differences (Sahin *et al.*, 2011). For a forensic application of epigenetics to discriminate between identical twins, a high throughput DNA methylation microarray was employed to select suitable CpG sites (Li *et al.*, 2011). Li *et al* tested blood DNA samples from 22 pairs of adult monozygotic twins, which were bisulphite treated and analysed with Illumina's Human Methylation 27K Beadchip, which allows for the measurement of methylation levels of more than 27,000 CpG sites. The results were very promising, revealing significant DNA methylation differences in a total of 377 CpG sites; however, further investigation is needed regarding the stability of CpG methylation and potential age-, tissue- or population-specific influence on DNA methylation.

1.2.4 Cause of death determination

The essential duties of the forensic pathologist include the investigation of the cause and manner of death, especially in sudden, unexpected or unwitnessed deaths. The main purpose of using post-mortem biochemistry and molecular biology is to examine the general pathophysiological changes involved in the death that cannot be established by morphological and functional changes of cells and organs (Maeda *et al.*, 2011; Zhao *et al.*, 2009).

Several studies aimed to use messenger RNA (mRNA) profiling in post-mortem human tissues to find clues on events that happened at the time point of death (Maeda *et al.*, 2010; Partemi *et al.*, 2010; Zhu *et al.*, 2008). The composition of different RNA transcripts within the cells of a tissue is not stable, but can change according to the specific needs of cells in response to environmental conditions. Thus, each type of death would activate different molecular mechanisms that could be detected through different gene expression patterns (Vennemann & Koppelkamm, 2010; Zhao *et al.*, 2009). However, mRNA instability could be an issue when working with human post-mortem tissues; gene transcripts are usually degraded by *in vivo* ubiquitous RNases shortly after death. This process might also be accelerated by several factors such as high humidity and temperature during post-mortem interval (PMI). Furthermore, it is

known that the integrity of RNA may differ between tissue types or different donors (Vennemann & Koppelkamm, 2010).

As mentioned earlier, DNA methylation is an epigenetic mechanism that changes gene expression. Recent studies have demonstrated that various diseases and cancer are associated with abnormal alterations in DNA methylation levels (Szyf, 2010); therefore, it would be interesting to determine whether DNA methylation status in the promoter regions of certain genes could change as a response to various acquired diathesis and/or causes of death. The methylation status of the promoter region of *p16* gene (also known as cyclin-dependent kinase inhibitor 2A, multiple tumour suppressor 1, CDKN2A) was investigated in blood of individuals who had been exposed to lead (Kovatsi *et al.*, 2010). One of the main symptoms of lead poisoning is the development of neurological disorders, closely related to DNA methylation changes (Fragou *et al.*, 2011). In this study, methylation levels were detected through methylation-specific PCR followed by thermal denaturation. *p16* methylation was indeed frequent and extensive among lead-overexposed individuals and dependant on blood lead concentration. Highly exposed individuals (blood Pb^{2+} concentration: 51-100 μg/dL) showed complete methylation, whereas those with low blood Pb^{2+} concentration (6-11 μg/dL) demonstrated partial methylation (Kovatsi *et al.*, 2010).

Moreover, the methylation status of circadian clock gene promoters has been examined using forensic autopsy specimens (Nakatome *et al.*, 2011). Circadian clock genes play a vital role in the formation of the biological clock and aberrant methylation of these genes contributes to several diseases such as lifestyle-related or heart diseases. Pathomorphological changes are difficult to be detected in cases of sudden unexpected death caused by cardiac arrhythmia or certain long-term medication. The majority of patients who are liable to sudden death lose the balance of a range of biological functions related to circadian clock genes (Nakatome *et al.*, 2011). The authors analysed nine circadian clock genes - period circadian protein homologs 1-3 (*Per1, Per2, Per3*), cryptochrome circadian clock 1 and 2 (*Cry1, Cry2*), timeless *(Tim),* casein kinase 1 *(Ck1e)* and the clock component *Bmal1* - by employing bisulphite treatment followed by methylation-specific PCR. Tissues were obtained from several cadavers with known or unknown cause of death (asphyxia, lethal arrhythmia, head injury, fire fatality). The methylation status of these promoter regions was found to vary

significantly between individuals and among tissues in the same individual. Even though it has previously reported that circadian dysregulation is relevant with drug addiction or alcohol abuse (Li *et al.*, 2010), this study failed to demonstrate a link between methylation pattern and cause of death (Nakatome *et al.*, 2011). As with the other studies there are some interesting links of methylation patterns and cause of death but it is too early to draw any definite conclusions.

1.2.5 Challenges and practical considerations

DNA-based testing is favoured in forensic analysis as DNA is considered to be among the most stable molecules present in cells. Assays based on the use of epigenetic markers that by default use the same starting material for downstream profiling have obvious advantages. However, caution is needed regarding the selection of CpG sites as well as the interpretation of the observed DNA methylation profiles.

The first step for a successful epigenetic assay is the identification of suitable DNA methylation markers. There are two different approaches in selecting CpG sites; the first involves screening well-known candidate regions such as CpG islands and gene promoters, while the second employs high-throughput analysis for the epigenetic biomarker discovery on a truly genome-wide scale (Bock, 2009).

Regarding the first approach and as discussed earlier in this chapter, it is generally believed that DNA methylation is associated with silencing of gene expression (Newell-Price *et al.*, 2000), perhaps by blocking the gene promoters at which transcription activators should bind. As discussed by Suzuki and Bird, evidence of this has been highlighted in studies showing that DNA methylation patterns near gene promoters considerably varies depending on the cell-type, with higher levels of methylation correlating with very low or no transcription (Suzuki & Bird, 2008). Therefore, DNA methylation is known to play a role in cell differentiation and adaptation to different cell environments, mainly by regulating the expression of selected genes. Depending on the forensic application investigating well-known candidate genomic regions would potentially reveal suitable epigenetic markers. On the other hand, a genome-wide DNA methylation analysis is very likely to uncover many new genomic regions that exhibit epigenetic variations, including those that are located outside gene promoters, which would be otherwise missed following the first approach. However, this increase in the scale of the search brings about major

bioinformatics and statistical challenges; however, these challenges can be resolved utilising computational methods. The proposed methods are only used for a first screening and identification of a pool of CpG sites; whereas the validation of the selected markers needs to be achieved using highly-targeted assays (Bock, 2009).

As mentioned earlier, DNA methylation as an epigenetic mechanism is responsible for a range of activities inside the cell, therefore external factors such as ageing, environmental stress, human diseases, cancer and cigarette smoking could interfere by altering DNA methylation patterns (Bocklandt *et al.*, 2011; Christensen *et al.*, 2009; Koch & Wagner, 2011; Wild & Flanagan, 2010). Thus, following verification of the specific DNA methylation pattern of a given CpG site, it is necessary to analyse a large number of DNA samples and ensure that this particular CpG site does not display high variation in its methylation levels (low inter-individual variation). Moreover, for each forensic application all aforementioned external factors need to be tested. For example, in order to propose that a particular CpG site demonstrates body-fluid specific DNA methylation pattern, it needs to be confirmed that it is not also age-dependent or influenced by environmental factors (Christensen *et al.*, 2009).

In cases where the levels of DNA methylation can be quantified, it is essential to choose CpG sites that clearly show a high degree of methylation difference; the larger this margin the better the discrimination achieved. For instance, in order to correlate a specific CpG site with a particular cause of death, such as drug abuse, it needs to demonstrate a minimum of 60-70% difference compared to the normal tissue. This would also prevent false estimation due to inter-individual variation.

After selecting the right candidate CpG sites, the assay has to be carefully designed so that it can be applicable for forensic analysis. Forensic specimens are often of low quantity and quality, so it is important that a proposed method is tested on low levels and/or degraded DNA. Both bisulphite conversion protocols and assays using methylation-sensitive restriction enzymes show high degrees of sensitivity (LaRue *et al.*, 2013; Madi *et al.*, 2012). In both cases, appropriate controls are necessary (bisulphite conversion controls and enzyme digestion controls respectively).

Other very important steps of assay design are the PCR amplification and primer design. When sodium bisulphite is employed, the primers need to be designed to

contain some normally converted non-CpG cytosines, so that only the bisulphite-converted DNA is amplified. Also, primer sequences should not contain any CpG sites, except for the cases where methylation-specific PCR is chosen. Differences in the DNA sequence between methylated and unmethylated amplicons could introduce preferential PCR amplification; therefore, DNA methylation controls with known DNA methylation levels can be used to normalise the data (Moskalev *et al.*, 2011; Warnecke *et al.*, 1997).

As for any forensic testing, more stringent validation criteria have to be met for each selected epigenetic marker. All markers have to show high sensitivity and specificity, as forensic samples are frequently of low quality and quantity, aged or degraded.

1.3 Conclusion

The field of epigenetics is still considered relatively new, having gathered scientists' attention only the past few decades. There are many fundamental questions that are yet to be answered, and even more so when applying epigenetic models in forensic investigations. As the epigenetic code is adjustable and dynamic, changing with age and environment, forensic scientists need to keep up with new discoveries in scientific research. The promise of forensic epigenetics demands not only overcoming challenges regarding the design and interpretation of epigenetic profiles, but also exploring innovative strategies that could cope with the nature of forensic specimens. More comprehensive analyses of Forensic Epigenetics would allow for further identification of suitable epigenetic markers, more conclusive links between cell differentiation and ageing to epigenetic effects, and would shed additional light on an array of forensic applications.

1.4 Aims

This investigation aims to explore novel uses of epigenetics in forensic science by analysing various tissue-specific or age-associated differentially methylated CpG sites. Since this study looked into two distinct forensic applications, the thesis has been divided into two parts.

Part 1 includes findings regarding the detection of body fluids and tissues; in order to meet these aims, the objectives were:

- To evaluate the performance of existing multiplex mRNA-based PCR assays including assessing the sensitivity, specificity and applicability of proposed tissue-specific mRNA markers (Chapter 4)

- To identify novel and validate existing tissue-specific differentially methylated CpG sites via the development of bisulphite Pyrosequencing® assays (Chapter 5)

Part 2 includes findings regarding the estimation of an individual's age from blood where aims were met by the following objectives:

- To investigate a set of age-associated CpG sites proposed in the literature via bisulphite Pyrosequencing® and regression analysis and build an age prediction model through Artificial Neural Networks (ANN) (Chapter 7)

- To develop a next-generation sequencing protocol that could accurately measure CpG methylation levels for use in age prediction (Chapter 8)

2 Methodology

This chapter will describe all methods and materials used throughout the whole course of this research. More specific experimental and assay design will be discussed in detail in each of the following results chapters.

2.1 Samples

Biological samples included in this study were collected either for the purpose of paternity or relationship testing by the fully accredited 'DNA analysis at King's' laboratory in the Department of Pharmacy and Forensic Science (UKAS ISO17025, UK Ministry of Justice) or as part of this study following obtaining full ethical approval by the Biomedical Sciences, Dentistry, Medicine and Natural & Mathematical Sciences Research Ethics Subcommittee (BDM RESC 13/14-30, Appendix I). Samples were collected and stored in accordance with the Human Tissue Act Code of Practice on Consent (HTA, 2014) and according to the Human Genetics Commission documents relating to Genetic Testing Services (2010). Full informed consent for the testing performed was obtained from the donors or their parents in case of under-aged individuals prior to collection. For certain joint European projects, the collaborating laboratory obtained ethical approval. Lastly, stains from casework were collected after successful DNA analysis was already completed and according to the laboratory's instructions.

Individuals had the choice to donate one or more of the following body fluids/tissues: whole blood, saliva, buccal cells in the form of a mouth swab, seminal fluid, vaginal secretion, menstrual blood, urine, skin and/or nasal fluid. Up to 20 ml of whole blood were collected by a trained phlebotomist in a sterile location designed for blood sampling and collection. All other body fluid samples were collected either using a cotton swab or a suitable receptacle by the participants in the privacy of their own homes. Information such as subject's gender, ethnicity and age were also recorded if possible. Each sample was anonymised soon after collection; therefore, no genetic information obtained can be linked back or relate to an individual. All samples were stored at 4 ^{o}C until extraction and at -20 ^{o}C for long-term storage.

2.2 DNA analysis

Genomic DNA was extracted and analysed either for the purpose of DNA profiling or to investigate DNA methylation patterns.

2.2.1 DNA extraction

Body fluid samples were either in a liquid form, deposited on a swab or on a surface (fabric, glass). Part of these swabs or stains was used to extract and purify DNA from cellular material. The method to be used for DNA extraction was dependent on the sample type. For buccal swabs or samples where high quantities of DNA were likely to be present (like whole blood), manually performed extraction using Chelex™ beads (Sigma) or automated DNA purification system using an EZ1 instrument (QIAGEN) was performed. In cases where body fluid samples were suspected of being of low quality or quantity, the QIAamp DNA Investigator kit (QIAGEN) was employed, as it was possible to adjust the final elution volume for more concentrated DNA samples.

2.2.1.1 Chelex™ beads (Sigma)

Chelex™ is a chelating resin in the form of small beads that are capable of binding polyvalent metal ions such as magnesium (Mg^{2+}) and removing them from the solution. The Chelex™ resin is composed of styrene divinylbenzene copolymers containing paired iminodiacetate ions, which act as chelating groups (Singer-Sam *et al.*, 1989). The presence of these beads facilitates the inactivation of potentially harmful nucleases that could degrade the DNA during boiling in low ionic strength solutions. However, the resulting DNA is single-stranded and possible contaminants might not be entirely removed. The procedures used in this study were adapted and adjusted from the original published protocol (Walsh *et al.*, 1991).

Lysis of the cellular material was achieved by firstly incubating the samples in 1 ml of DNA- and nuclease-free water (Severn Biotech) at room temperature for 20 minutes. Samples are in the form of either 4 µl/1.2 mm dried spot of whole blood or the tip (3 mm) of a mouth swab. Following incubation, the tubes were spun down at 14,000 rpm for 5 minutes using a centrifuge (Biofuge pico), after which all liquid was removed except for 20 µl at the bottom of the tube. Afterwards, 180 µl of 5% w/v Chelex™ 100 in H_2O suspension were added to the tubes, which were then incubated on a heated shaker (ThermoStat Plus) at 56 ºC and 1,000 rpm for 20 minutes. They were then

incubated at 100 °C for exactly 8 minutes and spun down at 14,000 rpm for 5 minutes. The piece of mouth swab was not removed from the tube throughout the whole procedure in order to avoid any DNA loss. DNA samples were now ready for use and stored at 4 °C for up to 6 months.

2.2.1.2 QIAamp DNA Investigator (QIAGEN)

The QIAmp DNA Investigator kit uses an entrenched technology for the isolation of both genomic and mitochondrial DNA from small sample volumes and stain sizes. The protocol combines the selective binding properties of a silica-based membrane with flexible elution volumes (20-100 µl). This ability makes this kit suitable for analysing a broad range of forensic samples. The procedure is simple, quick and yields high-quality DNA. The method mainly comprises four steps including lysis, DNA binding to the membrane, removal of contaminants through wash steps and finally elution of pure and concentrated DNA. The protocol can be adjusted depending on the type of starting material and more information can be found in the QIAamp DNA Investigator Handbook (Qiagen, 2010b).

In this study, a combination of the protocol for 'body fluid stains', 'surface swabs' and 'small volumes of blood or saliva' was employed. In more detail, 50-100 µl of a body fluid, a whole stained cotton swab or a casework stain were placed into a labelled 1.5 ml tube. 100 to 400 µl of tissue lysis buffer (Buffer ATL) were then added together with 20 µl of proteinase K (both provided in the kit). If processing semen swabs, 20 µl of 1 M dithiothreitol ($C_4H_{10}O_2S_2$, DTT) (Thermo Scientific) were also added to the samples, as DTT is needed to break the sulphur bonds of various proteins on the surface of spermatozoa. The solutions were then incubated on a heated shaker at 56 °C and 1,000 rpm for at least 1 hour. When processing casework samples, the solutions were incubated for up to 3 hours to ensure sufficient lysis.

Following the incubation, another 300 to 600 µl of lysis buffer (Buffer AL) was added and the solutions were vortexed for at least 15 seconds. The tubes were again placed on a heated shaker at 70 °C and 1,000 rpm for 10 minutes. The samples were briefly centrifuged to remove any remaining drops from the inside of the lids. To enhance the binding of DNA to the silica membrane, 50 to 300 µl of absolute ethanol (96-100%) (VWA Chemicals) were added and the solutions were thoroughly mixed by vortexing for 15 seconds. The entire lysate was then transferred from the tubes to the QIAamp

MinElute columns, which are stored at 4 °C. The columns were centrifuged at 8,000 rpm for 1 minute and the flow through was discarded since DNA is expected to bind to the membrane. The first washing step was performed by adding 500 µl of Buffer AW1 followed by centrifugation at 8,000 rpm for 1 minute. The flow through was then discarded again and the process was repeated with both 700 µl Buffer AW2 and 700 µl of absolute ethanol. Both washing buffers were included in the kit.

The columns were then placed into new collection tubes and were centrifuged at full speed (14,000 rpm) for 3 minutes to completely dry the membrane. This step is necessary since ethanol carryover may interfere with downstream analysis. To ensure successful ethanol removal, the columns were also incubated with the lid open either at room temperature for 10 minutes or at 56 °C for 3 minutes. They were then placed into new, labelled 1.5 ml collection tubed and were ready for the elution step. Finally, 20-100 µl of elution buffer or nuclease-free water were added to the centre of the silica membrane and samples were incubated at room temperature for 1 minute in order to increase DNA yield. DNA was eluted after centrifuging the samples at full speed (14,000 rpm) for 1 minute. The last step was usually repeated by loading the eluate back to the membrane for maximum yield. DNA samples were stored at -20 °C.

2.2.1.3 BioRobotEZ1® DNA extraction (QIAGEN)

The EZ1® instrument offers reproducibly automated DNA extraction and purification employing a magnetic-particle technology. This method combines the speed and efficiency of silica-based DNA purification together with the expedient handling of magnetic beads. Firstly, DNA is isolated from the lysates during its binding to the silica surface of the particles in the presence of a chaotropic salt. Applying a magnet, the particles are separated from the solution and DNA can be efficiently eluted after it has been washed. This process is illustrated in Figure 2-1.

The protocols for DNA extraction from mouth swabs and whole blood were slightly different; therefore, there is a specific kit for each sample type. For blood samples, the EZ1® DNA blood kit together with the EZ1® DNA blood card were used; for mouth swabs, the EZ1® DNA tissue kit together with the EZ1® DNA tissue card were employed. Manufacturer's instructions can be found in the EZ1® handbooks (Qiagen, 2009b; Qiagen, 2011). Briefly, up to 6 samples can be extracted at the same time and the amount of starting blood material could vary, but usually 50-200 µl was used. The

methods are very similar; however for the mouth swabs there is an initial incubation step that facilitates cell lysis not present in the extraction protocol for blood. This preliminary step included placing the end 3 mm of the buccal swabs into a 2 ml sample tubes containing 190 µl of lysis buffer and 10 µl of proteinase K (both supplied with the kit). The solutions were incubated on a heating shaker at 56 °C and 1,000 rpm for 20 minutes.

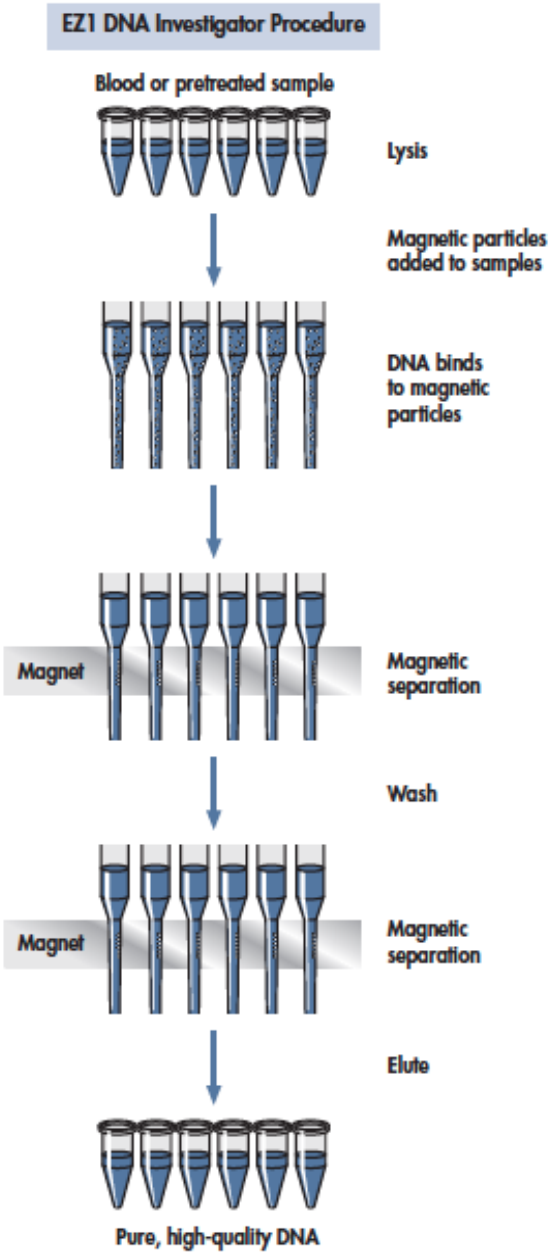


Figure 2-1. Principle and procedure of DNA extraction technology using EZ1® instruments (Qiagen, 2009b)

This diagram shows the magnetic bead technology employed in the BioRobot EZ1® protocols including cell lysis, DNA binding to the magnetic particles, magnetic particle removal, sample washes and elution of pure, high-quality DNA.

The blood or the pre-treated buccal cells were then put in the correct place and the appropriate card was inserted into the EZ1® robot. Following the on-screen instructions, reagent cartridges, tip holders with tips and 1.5 ml elution tubes (all supplied with the kit) were also placed in the specified positions. The elution volume could be selected depending on the starting volume (50-250 µl). The whole process lasts for about 20 minutes and purified DNA samples were stored at -20 °C until analysis. There is a chance that magnetic particles might still remain in the DNA solution; however, this does not affect downstream analysis.

2.2.2 DNA quantification

It was necessary to determine the concentration of DNA samples before proceeding to further analysis. A method that can accurately evaluate the quality and quantity in a DNA sample in a human-specific manner was preferred since it is very likely that some of the tissues involved in this research, such as saliva and vaginal fluid, are highly contaminated with bacterial DNA.

2.2.2.1 Quantifiler® Human DNA Quantification (Applied Biosystems)

This real-time PCR-based system is designed for the accurate quantification of human DNA present in a solution and has been widely used in the forensic field. DNA from any human tissue source can be quantified and possible PCR inhibitors can be also detected with the use of a pre-formulated internal PCR control (IPC), which monitors the amplification success of each sample. The IPC is a synthetic template DNA not found in nature that is co-amplified in the reaction. The chemistry used is compatible with commonly used extraction technologies and currently has the widest dynamic range of DNA detection (0.23 to 50 ng/µl) allowing for true quantification of low-level samples.

The Quantifiler® Human DNA Quantification kit takes advantage of the TaqMan® technology, which is used to detect and amplify a 62 bp long sequence in a non-translated region (introns) of the human telomerase reverse transcriptase gene (*hTERT*) at the chromosomal location 5p15.33. The method involves a TaqMan® probe which contains a reporter dye (FAM™ or VIC™) linked to its 5' end, a minor groove binder (MGB) at its 3' end together with a non-fluorescent quencher (NFQ). During PCR, the TaqMan® MGB probe specifically anneals to a complementary

sequence between the forward and the reverse primer sites. When the probe is intact, the proximity of the reporter dye to the quencher dye leads to repression of the reporter fluorescence. However, cleavage by the AmpliTaq Gold® DNA polymerase during polymerisation results in separation of the reporter dye from the quencher dye, which causes fluorescence [Figure 2-2]. The increase in fluorescent signal takes place only when the target sequence is complementary to the probe and is amplified during PCR. Therefore, it is exponential and proportional to the amount of starting DNA present in the sample and can be easily determined with the aid of DNA standards of known concentration. The ABI PRISM® 7000 Sequence Detection System was used to monitor the fluorescent signal. More details on the technology can be found in the manufacturer's recommendations (ABI, 2014).

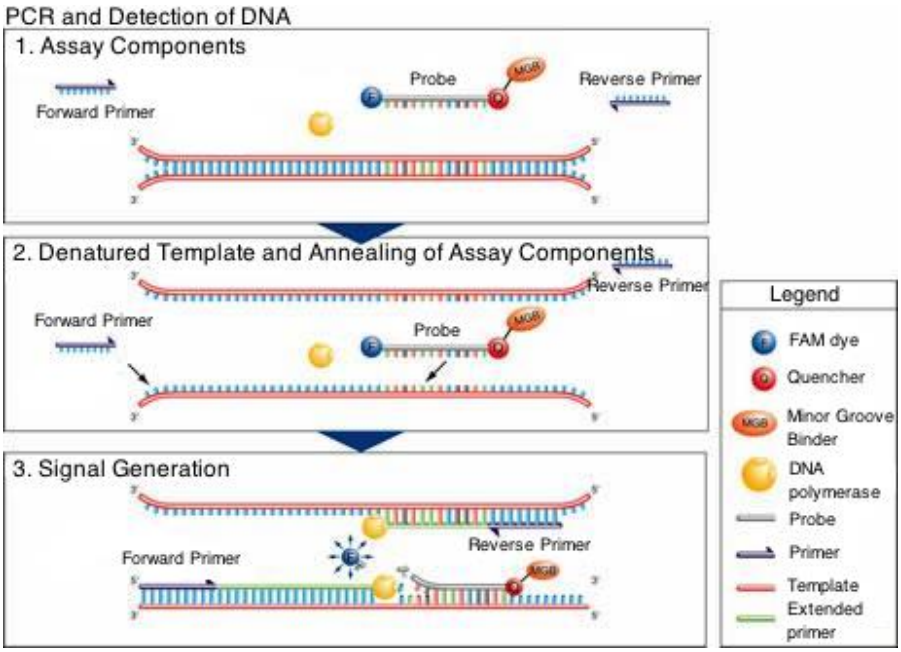


Figure 2-2. TaqMan® technology used in the Quantifiler® Human DNA Quantification kit (ABI, 2014)

The TaqMan® MGB probe binds specifically to a complementary sequence between the forward and the reverse primer sites. When the probe is intact, the proximity of the reporter dye to the quencher dye leads to suppression of the reporter fluorescence. However, when the AmpliTaq Gold® DNA polymerase synthesises the DNA during PCR it removes the reporter dye from the quencher dye resulting in fluorescence.

Firstly, the DNA standards were prepared by mixing the standard stock solution (200 ng/µl) that is provided with the kit with nuclease-free water as shown in Table 2-1. It is important that the stock solution is thoroughly defrosted and that all standards are vortexed very well during preparation. Quantification was carried out in 25 µl reactions and in duplicate for both standards and samples; negative control (H_2O) was

also included in each experiment. For each reaction, 10.5 µl of the Quantifiler® Human Primer mix were added to 12.5 µl of Quantifiler® PCR reaction mix (both supplied with the kit) along with 2 µl of DNA standard, DNA sample or water in a 96-well plate. To avoid contamination, only filter tips were used when handling the samples. The plate was sealed, vortexed and briefly centrifuged to remove any bubbles. The plate was then loaded in the ABI PRISM® 7000 instrument for analysis. The thermal cycling conditions used were 95 °C for 10 minutes followed by 40 cycles of 95 °C for 15 seconds and 60 °C for 1 minute. The whole method lasts for about 2 hours. After the run is complete, it is analysed according to the analysis parameters suggested by the manufacturer. The R^2 value of the standard curve needs to always be >0.985 with a slope of 2.7<s<3.3. In case of a lower R^2 value, up to two standards could be omitted from the standard curve; if it was still <0.985, the run was repeated.

Table 2-1. DNA standards dilution series

Standard	Concentration (ng/µl)	Dilutions	Dilution factor
Std. 1	50.000	10 µl stock + 30 µl H_2O	4X
Std. 2	16.700	10 µl Std. 1 + 20 µl H_2O	3X
Std. 3	5.560	10 µl Std. 2 + 20 µl H_2O	3X
Std. 4	1.850	10 µl Std. 3 + 20 µl H_2O	3X
Std. 5	0.620	10 µl Std. 4 + 20 µl H_2O	3X
Std. 6	0.210	10 µl Std. 5 + 20 µl H_2O	3X
Std. 7	0.068	10 µl Std. 6 + 20 µl H_2O	3X
Std. 8	0.023	10 µl Std. 7 + 20 µl H_2O	3X

2.2.3 DNA treatment with sodium bisulphite

In order to convert differences in DNA methylation to differences in DNA sequence, DNA samples are treated with sodium bisulphite, which converts unmethylated cytosines into uracil, while the methylated ones remain unchanged. Once converted, the methylation profile of DNA regarding any locus can be determined by PCR amplification followed by DNA sequencing. In this research, three commercially available bisulphite conversion methods were tested for their efficacy and accuracy, including the Epitect® Bisulphite kit (QIAGEN), the EZ DNA Methylation™ kit (ZymoResearch) and the MethylEdge™ Bisulphite Conversion system (Promega). All kits comprise a few simple steps [Figure 2-3].

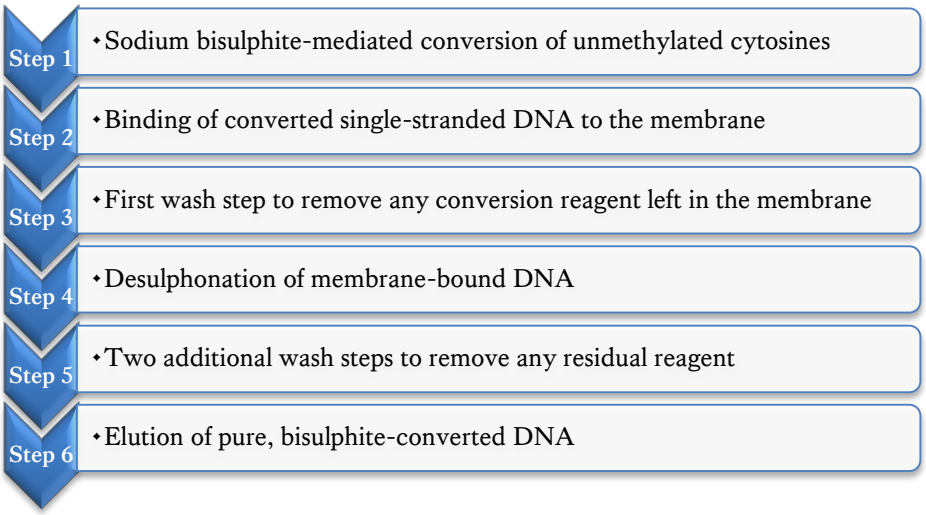


Figure 2-3. Main steps of bisulphite conversion protocols

2.2.3.1 Epitect® Bisulphite (QIAGEN)

The EpiTect® Bisulphite kit provides a fast and streamlined 6-hour procedure for complete conversion and purification of as little as 1 ng of starting DNA material. The use of a DNA Protect Buffer (supplied with the kit) prevents DNA fragmentation and loss due to high temperatures and low pH values as it includes a control pH-indicator dye that allows for verification of the correct pH for cytosine conversion. Additionally, the kit allows for the analysis of difficult samples such as small amounts of fragmented DNA with adjusted protocols including an additional step that enhances binding of DNA and considers starting DNA amounts of less than 500 pg (Qiagen, 2009a).

The standard protocol allows for complete conversion of 1 ng – 2 µg DNA in a volume of up to 20 µl. Each aliquot of Bisulphite mix was diluted with 800 µl of RNase-free water (provided in the kit) and vortexed for 5 minutes. Each bisulphite reaction consisted of 20 µl of DNA solution (DNA was diluted with water if necessary), 85 µl of dissolved Bisulphite mix and 35 µl of DNA Protect Buffer for a total volume of 140 µl. The bisulphite DNA conversion was performed using a Veriti® thermal cycler (Life Technologies) using the program shown in Table 2-2. Following this 5-hour long process, the clean-up of bisulphite converted DNA was carried out.

The complete bisulphite reactions were transferred to new 1.5 ml tubes and 560 µl of freshly prepared Buffer BL (containing 10 µg/ml carrier RNA if <100 ng DNA were used) were added. The solutions were mixed by vortexing and transferred to the EpiTect® spin columns, which were then centrifuged at 14,000 rpm for 1 minute. The

flow through was discarded and the first wash step was performed by adding 500 µl of wash buffer (Buffer BW) and centrifuging at 14,000 rpm for 1 minute. Next, 500 µl of desulphonation buffer (Buffer BD) was added to each column and left at room temperature for 15 minutes. The spin columns were then centrifuged, the flow through was discarded and two more washing step using 500 µl of Buffer BW were carried out. To enable removal or evaporation of any remaining liquid, the columns were centrifuged at maximum speed for 1 minute and also incubated in a heating block at 56 °C for 5 minutes. The bisulphite-converted DNA strands were now ready to be eluted by dispensing 20-40 µl of elution buffer (Buffer EB) onto the centre of each membrane and centrifuging at 12,000 rpm for 1 minute. To increase the yield of DNA in the eluate, the columns were incubated at room temperature for 1 minute prior to the centrifugation step. The purified DNA samples could be stored at 4 °C for up to 24 hours or at -20 °C for long-term storage.

Table 2-2. Bisulphite conversion thermal cycler conditions (Epitect®, QIAGEN)

Step	Time	Temperature
Denaturation	5 min	95 °C
Incubation	25 min	60 °C
Denaturation	5 min	95 °C
Incubation	85 min (1 h 25 min)	60 °C
Denaturation	5 min	95 °C
Incubation	175 min (2 h 55 min)	60 °C
Hold	∞	20 °C

2.2.3.2 EZ DNA methylation™ (ZymoResearch)

In a very similar way, the EZ DNA methylation™ kit is based on a three-step reaction where unmethylated cytosine is converted to uracil. In-column desulphonation technology is designed to reduce destructive precipitations and at the same time eliminate any DNA fragmentation. The kit specifies that the method can be used for 500 pg to 2 µg starting DNA material, but for optimal results, the input DNA should be from 200 to 500 ng, since the final DNA recovery has been calculated to be around 80% (ZymoResearch, 2013a).

For the purpose of this research, modified conversion conditions were applied as explained in this paragraph. All reagents and buffers are supplied with the kit. The CT

conversion reagent was a solid mixture, therefore was diluted with 750 µl of distilled water and 185 µl of dilution buffer and left at room temperature for at least 20 minutes with frequent vortexing to ensure adequate dissolving. 7.5 µl of dilution buffer was also added to the DNA samples and the final volume was adjusted to 50 µl with water. The solutions were then incubated at 42 °C for 30 minutes followed by the addition of 97.5 µl of prepared conversion reagent. The samples were left at 50 °C for 12-16 hours (usually overnight) before the clean-up took place. 400 µl of binding buffer was added to the Zymo-Spin™ IC columns together with the samples and the mix was inverted several times. In order for the bisulphite-converted DNA to bind onto the membrane, the samples were centrifuged at 14,000 rpm for 30 seconds and the flow through was discarded. A first washing step was performed with 100 µl of wash buffer and the desulphonation procedure took place using 200 µl of desulphonation buffer at room temperature for 20 minutes. Samples were centrifuged at full speed for 30 seconds and then two extra washing steps were carried out as before. Bisulphite-treated DNA samples were eluted using 10-20 µl of elution buffer and were stored at -20 °C.

2.2.3.3 MethylEdge™ bisulphite conversion (Promega)

As shown above, a bisulphite conversion protocol can last between 6-20 hours, therefore it would be very beneficial to have a method that is quick but offers complete conversion at the same time. The MethylEdge™ Bisulphite Conversion system by Promega offers a novel and efficient method that can be completed in less than 2 hours and it can convert 100 pg - 2 µg of DNA (optimal range 200-500 ng). It is suggested that prolonged incubation times under harsh conditions can degrade DNA molecules emphasising the advantages of a quicker protocol. The protocol was followed as it is explained in the manufacturer's technical manual (Promega, 2013).

20 µl of aliquots of extracted DNA (volume adjusted with distilled water) were mixed with 130 µl of ready-to-use ME Conversion reagent and the solutions were mixed by pipetting. The DNA solutions were placed in a thermocycler and incubated at 98 °C for 8 minutes and 54 °C for 60 minutes. The clean-up process was very similar with the previously described methods, including adding 600 µl of binding buffer, one wash step (100 µl wash buffer), adding 200 µl desulphonation buffer and incubating at room temperature for 15 minutes, two additional wash steps (200 µl wash buffer) and finally,

eluting using 10-20 µl of elution buffer. Samples were stored at 4 °C for use within one week and at -20 °C for long-term use.

2.2.3.4 DNA methylation standards (QIAGEN, ZymoResearch, EpigenDx)

In DNA methylation analysis, it is important that appropriate methylation controls are used. These are DNA standards of known methylation status (0-100%) that are employed to assess assay performance. They not only confirm that the bisulphite conversion is complete, but also evaluate the accuracy and reproducibility of the chosen methylation quantification procedure. For this study, several commercially available DNA methylation standards were used: the EpiTect® PCR Control DNA kit (QIAGEN), the Human Methylated and Non-Methylated DNA set (ZymoResearch) and the Human Premixed Calibration Methylation standards (EpiGenDx).The EpiTect® PCR Control DNA kit contains three different types of human DNA – the methylated bisulphite-converted, the unmethylated bisulphite-converted and the unmethylated non-converted control. All solutions have a concentration of 10 ng/µl in Buffer EB (QIAGEN) and a total volume of 100 µl. Complete *in vitro* methylation of the control DNA was achieved using the enzyme CpG methyltransferase, M.SssI and bisulphite conversion was completed using the EpiTect® Bisulphite kit.

The Human Methylated and Non-Methylated DNA set consists of purified, non-methylated and methylated human DKO DNA (5 µg/20 µl) as well as specifically designed control primers (20 µl) that are intended to assess the effectiveness of bisulphite conversion. They are designed to amplify non-methylated, methylated, and varied methylation copies of the death-associated protein kinase 1 gene (*DAPK1*) following bisulphite treatment. The human non-methylated control is isolated from cells that contain genetic knockouts of both DNA methyltransferases (DNMT1 and DNMT3b); therefore it has a low level of DNA methylation in all CpG sites across the genome. On the other hand, the methylated control DNA has been enzymatically methylated at all CpG sites with the M.SssI methyltransferase. Finally, the Human Premixed Calibration Methylation standards are manufactured known methylation DNA controls (0%, 5%, 10%, 25%, 50%, 75%, 100%) (50 ng/µl). They are made as a result of mixing a high-methylated DNA (>85%) and a low-methylated DNA (<5%) (both enzymatically made) in the right ratio.

2.2.4 DNA amplification

PCR is one of the most commonly used techniques in the field of molecular biology and genetics. It revolutionised the way we analyse DNA variation since it allows for the amplification of small DNA regions in an exponential manner. Therefore, it facilitates the analysis of the areas of interest from minute amounts of starting material, which is particularly important when the samples are aged or degraded. The process is simple and resembles the DNA replication that occurs inside the cells.

Regardless of the application, genome or area of interest, a standard PCR is composed of a few fundamental components. Firstly, the template DNA is denatured and the two strands are separated. Following denaturation, primers, which are small synthetic DNA oligonucleotides (18-35 bp long), recognise and specifically bind to DNA regions surrounding the area of interest. Since DNA is double-stranded, two primers - forward and reverse - are needed to copy the DNA, one strand each. The primer/DNA complex then acts as a starting point for the DNA polymerase. The enzyme used in PCR is the *Taq* polymerase that comes from the bacterium *Thermus aquaticus*. This type of bacteria lives in hot springs, therefore, the enzyme is thermostable and can 'survive' the hot temperatures (95 ^{o}C) needed for DNA denaturation. Using DNA as a template, the enzyme then adds deoxyribonucleotide triphosphates (dNTPs) in a complementary manner and extends the DNA strand. The nucleotides, also known as building blocks, are needed to generate the DNA copies of the area that the primers span. However, as with all enzymes, DNA polymerase can reach its maximum activity only in certain environment. Thus, a reaction buffer is needed to control the pH and the salt concentration of the solution. Also, the divalent cation magnesium chloride ($MgCl_2$) is needed as it promotes DNA annealing and serves as a co-factor for DNA polymerase activity. Once the first copy has been created, the DNA strands are denatured again and the whole procedure is repeated.

Consequently, DNA amplification happens during three main steps, denaturation, annealing and extension. These steps require different temperatures; therefore a thermal cycler is employed. To start with, the denaturation step involves a high temperature of 94-95 ^{o}C so that the two strands completely break apart. Secondly, the annealing process is primer-specific but usually temperatures between 45-65 ^{o}C are required; the primer length and sequence are the main factors that determine the

annealing temperature. Lastly, the extension step where DNA polymerase adds the nucleotides is typically performed at 72 °C, the enzyme's optimum temperature. By repeating these steps again and again, a small number of DNA copies can result in millions after 20 or 30 cycles. A final extension step is also added following the cycles to ensure complete adenylation of the PCR products. *Taq* polymerase tends to add an additional base (adenosine) to the end of each product; therefore incomplete adenylation has to be avoided as it results in products that differ by a base (Butler, 2012). In conclusion, using the method of PCR, researchers are able to amplify any part of a genome, either human or bacterial, or bisulphite-converted DNA.

2.2.4.1 Primer design

As expected, primers are very important for the success of a PCR reaction, and careful design is necessary for a successful amplification with no non-specific signal. In this study, primers had to be designed in order to amplify bisulphite-treated DNA, therefore specific design parameters needed to be adjusted as bisulphite PCR is generally known to be of low efficiency. Mis-priming events as well as non-specific amplification are more likely to occur due to the T-richness of the bisulphite-treated DNA sequences (all non-CpG cytosines have been converted to thymines). There are currently various primer design software specifically designed for methylation analysis; however, the online-tool BiSearch was considered as most suitable and selected for assay design (Tusnady *et al.*, 2005). This software has the distinctive property of applying an algorithm and therefore, suggests potential mis-priming sites for the designed primer pair as well as determining possible non-specific amplicons. It can be applied in both treated and untreated genomes, however it is suggested that its new PCR primer design strategy increases the amplification efficiency of treated templates (Tusnady *et al.*, 2005).

BiSearch allowed for flexible primer design as parameters like primer length, melting temperature and PCR length could be manually adjusted. In general, PCR primers needed to flank the CpGs of interest but contain no CpG site in their sequence. They were designed to be 18-30 bp long; longer primers were preferred since they would be more specific, however when amplifying CpG-rich DNA regions designing primers longer than 18 bp was sometimes impossible. The PCR products were aimed to be as short as possible (100-300 bp) so that the designed assays could be applied in degraded

or casework samples. The optimal GC content of the primers (for bisulphite-treated DNA) was chosen to be 30% and the melting temperature (T_m) between 45 °C and 65 °C. Higher temperatures would result in minimising the presence of non-specific PCR products. Also, the forward and reverse primers should share a similar melting temperature (maximum T_m difference ~3 °C), as the PCR optimisation could be difficult otherwise. Furthermore, the software took into account melting differences between the methylated and unmethylated 'versions' of the PCR product. A maximum T_m difference of 2.5 °C was applied so that potential amplification bias was eliminated. A summary of these parameters is presented in Table 2-3 and an example of bisulphite PCR assay design in illustrated in Appendix II.

Table 2-3. General primer design parameters applied in BiSearch software (Tusnady *et al.*, 2005)

Parameters	Minimum	Optimal	Maximum
Primer length (bp)	17	**23**	30
Primer GC content (original) (%)	40	**50**	60
Primer GC content (bisulphite) (%)	0	**30**	60
Primer melting temperature (T_m) (°C)	45	**55**	65
T_m difference between primers (°C)	0	**0**	5
PCR length (bp)	100	**200**	300

The secondary structure of primers is also important; therefore, the design parameters regarding primer self- or pair- annealing had to be strictly controlled. Wherever possible, long runs of the same base (-TTTTTTT-) were avoided as this causes the primer to twist or curve. Self end-annealing was also considered in order to avoid the formation of hairpin structures. To further check for interaction between primers or between primers and the PCR product, an additional program called AutoDimer was also employed (Vallone & Butler, 2004).

When PCR products were to be sequenced by Pyrosequencing®, the reverse primers had to be biotin-labelled (5' end). Also, the design of sequencing primers was much easier since they just needed to bind to the biotinylated strand within 2-5 bases before the CpG of interest. Sequencing primers were chosen to be at least 22 bp long. Since the sequencing reaction is run at 28 °C, it was especially important to check the primers for self-annealing, in particular at the 3' end.

Moreover, as mentioned above BiSearch allowed for screening the entire bisulphite transformed genome and detection of possible mis-priming events of both the forward and reverse primers. Possible non-specific products were avoided but in cases where the number of potential primer pairs suggested by the software was limited, they were chosen not to be of the same length with the desired amplicon so that appropriate PCR optimisation could potentially eliminate them. Moreover, it is important that the primers bind specifically to bisulphite-converted DNA and not genomic DNA, therefore primers with no converted non-CpG cytosines were excluded. To check for possible mis-priming events to genomic DNA, the NCBI BLAST algorithm (www.ncbi.nlm.nih.gov/blast) was also used. All primers were ordered from biomers.net and were HPLC-purified.

2.2.4.2 Bisulphite PCR optimisation

PCR optimisation is the procedure by which the PCR product yield is maximised through applying optimal PCR reaction conditions. The primer design software usually provides suggested PCR conditions, however these are not always the most favourable. Optimising the PCR reaction can be achieved by adjusting not only the PCR reaction parameters such as primer or magnesium chloride ($MgCl_2$) concentration but also the thermal cycling conditions like annealing temperature and duration of each different cycle step. In general, regarding $MgCl_2$, optimisation can be challenging. Increased concentration usually causes higher incorporation rate and efficiency for the DNA polymerase but can also make the enzyme a little less specific increasing the risk of incorporation errors. Also, it can further stabilise dimer formation allowing for mis-priming or secondary structures in the DNA template. On the other hand, reduced concentration can make the PCR reaction more specific, but at the same time less efficient producing lower yields.

2.2.4.3 PyroMark PCR (QIAGEN)

The PyroMark PCR kit is specifically designed and optimised for Pyrosequencing® analysis that promises highly specific and unbiased amplification of template DNA for any Pyrosequencing® application such as methylation analysis (Qiagen, 2009d). The unique master mix solution guarantees that the primers only bind their specific targets with no mis-priming possibilities because of the balanced salt combination (KCl and $(NH_4)_2SO_4$). It also inhibits the build-up of biotinylated reverse primer excess enabling

reliable sequencing results. Also, the DNA polymerase used in the kit is a HotStarTaq, which is a modified version of the recombinant 94 kDa *Taq* DNA polymerase. The kit is provided with a novel PCR additive called Q Solution that facilitates amplification of difficult templates such as bisulphite-converted DNA sequences through modifying how DNA melts. Lastly, the kit is supplied with CoralLoad Concentrate, which not only assists in producing a high yield of amplified DNA, but also allows for direct loading onto an agarose gel without loading buffer because of its bright colour.

The 25 µl PCR reactions were set up using either thin-walled microtubes (Anachem) or 96-well plates (Starlab) and consisted of 12.5 µl of PyroMark PCR Master Mix (2x), 2.5 µl CoralLoad Concentrate (10X), 1 µl of each primer (10 µM) (forward and biotinylated reverse), 5 µl of Q Solution (5X), 2 µl of DNase-free water and 1 µl of DNA template. No additional Mg^{2+} was added; therefore the final concentration was 1.5 mM. The samples were then placed in a Veriti® thermal cycler (Life Technologies), which was programmed as follows: 95 °C for 15 minutes followed by 45 cycles of 94 °C for 30 seconds, 55 °C for 30 seconds and 72 °C for 30 seconds as well as a final extension step of 72 °C for 10 minutes. PCR products were stored at 4 °C prior to sequencing analysis.

2.2.4.4 ZymoTaq™ Premix (ZymoResearch)

Similarly, the ZymoTaq™ Premix is optimised for the amplification of bisulphite-treated DNA for methylation detection (ZymoResearch, 2013b). A hot-start DNA polymerase included in the 'master mix' allows robust product formation by reducing the presence of non-specific products. The set-up of the PCR reactions is very simple as all required components are included in the mix apart from the primers and DNA template. In this study, 25 µl reactions were proposed including 12.5 µl of ZymoTaq™ Premix (2X), 1 µl of each primer (10 µM) for a final concentration of 0.4 µM, 1 µl of 25 mM $MgCl_2$ solution for a final concentration of 2.75 mM (since the ZymoTaq™ Premix also contains 1.75 mM $MgCl_2$), 1 µl of DNA template and 8.5 µl of DNase-free water. The suggested PCR cycling conditions were: 95 °C for 10 minutes followed by 30-45 cycles of 94 °C for 30 seconds, T_m (50-60 °C) for 30-40 seconds and 72 °C for 30-60 seconds as well as a final extension step of 72 °C for 7-15 minutes. PCR products were stored at 4 °C until sequencing analysis.

2.2.4.5 PowerPlex® ESI 16 (Promega)

In this research, DNA profiling was performed in cases where the unique DNA profile of an individual needed to be obtained or as a proof that the body fluid samples originated only from one source. For example, it was particularly important to avoid contamination of seminal fluid in vaginal secretion swabs. For this purpose, the PowerPlex® ESI 16 system was chosen which allows for the co-amplification and four-colour detection of 16 loci (15 STRs and amelogenin), including D22S1045, D2S1338, D19S433, D3S1358, Amelogenin, D2S441, D10S1248, D1S1656, D18S51, D16S539, D12S391, D21S11, vWA, TH01, FGA and D8S1179 (Promega, 2014). All materials necessary are included in the kit, including the hot-start *Taq* DNA polymerase, which is a built-in component of the PowerPlex® ESI 5X Master mix. In this study, an adjusted protocol with reduced volumes was employed. Each PCR reaction consisted of 2 µl of PowerPlex® ESI 5X Master mix, 1 µl of PowerPlex® ESI 16 10X Primer Pair mix, 1 µl of template DNA (~1 ng) and 6 µl of nuclease-free water up to a final volume of 10 µl. For every experiment, a positive (Control DNA 007, Applied Biosystems) and a negative control (distilled water) were also included for assay assessment. The samples were then sealed and placed in a Veriti® thermal cycler (Life Technologies) with the following program: 96 °C for 2 minutes followed by 28 cycles of 94 °C for 30 seconds, 59 °C for 2 minutes and 72 °C for 90 seconds as well as a final extension step of 60 °C for 45 minutes. PCR products were stored at 4 °C until fragment analysis.

2.2.5 DNA fragment analysis

The amplified PCR products could be detected with various techniques. In the case of bisulphite PCRs, a confirmation of successful amplification was needed through agarose gel electrophoresis. Possible non-specific amplicons could also be detected. On the other hand, in cases of DNA profiling, the STR amplicons could be analysed by capillary electrophoresis.

2.2.5.1 Agarose gel electrophoresis

Agarose gel electrophoresis is a very common and effective technique for the separation and detection of DNA fragments of varying sizes (50 bp – 25 kb) (Lee *et al.*, 2012b). During the gel formation, agarose polymers create a network of pores of

various sizes, which determine a gel's properties. DNA samples are loaded in pre-cast wells in the gel and a current is applied. Since the phosphate backbone of DNA is negatively charged, the DNA fragments migrate to the positively charged anode in an electric field. Due to DNA's distinctive mass-to-charge ratio, the fragments are also separated by size in a way that smaller molecules travel faster and more easily through the agarose gel pores. Dyes such as ethidium bromide or GelRed® interact with DNA and stain it so that it can be seen under UV light. The band intensity on the gel is proportional to the DNA amount, so agarose gel can also be used for quantification (Lee *et al.*, 2012b).

In this study, 3.4 g of analytical grade agarose powder (Promega) were mixed with 170 ml of 0.5X TBE buffer (Sigma) and 17 µl of GelRed® stain (VWR) to obtain a 2% agarose gel. 5 µl of each PCR product were mixed with 1 µl of loading buffer (Bioline) and were then loaded into a well on the gel. The gel was left to run for 1 hour at 100 V in 0.5X TBE buffer. A size ladder of various fragments (25-500 bp) (Hyperladder™ 25bp, Bioline) was also applied as a size reference. Once the electrophoresis was complete, the gel was visualised under UV light in an AlphaImager® gel documentation system (Alpha Innotech Corporation). The presence of a bright, dense single band of the expected size suggests that a PCR was successful [Figure 2-4].

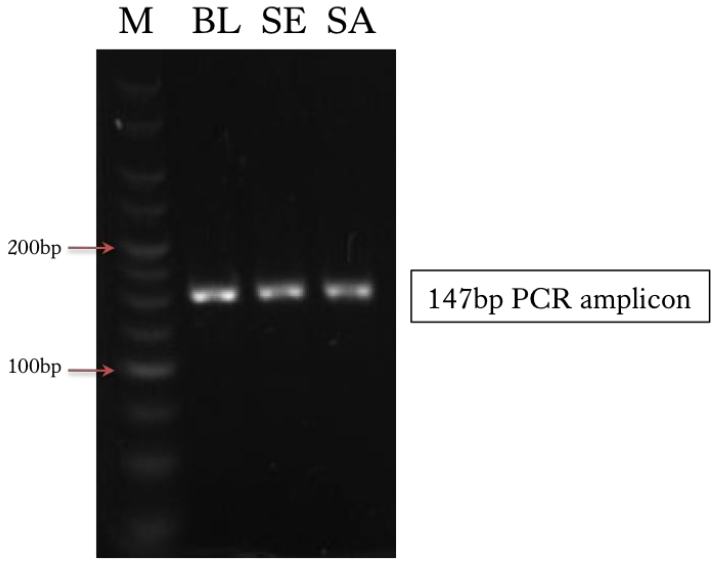


Figure 2-4. Example of agarose gel electrophoresis

This is a picture obtained by exposing a stained agarose gel under UV light. The first column corresponds to the DNA marker (M) consisting of various DNA fragments of known length (25-500 bp). Columns 2-4 contain amplified DNA from various tissue sources (Blood-BL, Semen-SE, Saliva-SA) of a particular genomic region, which is 147 bp long. As shown, all bands give a quite strong signal with no non-specific PCR products present.

2.2.5.2 Capillary electrophoresis (CE)

As mentioned above, the STR amplicons are each labelled with a fluorescent dye that allows for their detection and separation. The determination of the amplicon size is also very important in DNA profiling as it corresponds to the number of repeats in each analysed STR. STR amplicons are now usually separated by capillary electrophoresis, which is much more sensitive than the agarose gel electrophoresis. PCR products can be separated firstly due to DNA's negative charge, secondly due to their size (shorter PCR products are eluted first) and lastly, by their different coloured dye. In this project, an ABI-Prism-3130xl genetic analyser (Life Technologies) was used. The instrument has the ability to analyse up to 96 samples by running each sample through a polymer-filled (POP-7™, Life Technologies) capillary; it has 16 capillaries so it can process 16 samples at a time. PCR products of up to 500 bp long can be separated within approximately 40 minutes. A laser excites the different dyes, whilst a charge coupled device (CCD) camera detects the emitted fluorescence.

DNA samples are prepared in 10 µl reactions as follows: 1 µl of PCR product is mixed together with 8.6 µl of ultra-pure deionised formamide Hi-Di™ (Life Technologies) and 0.4 µl of GeneScan™ 500 LIZ® internal size standard (Life Technologies) in a 96-well plate (Life Technologies). As a size reference, 1 µl of PowerPlex ESI 16 Allelic Ladder Mix (Promega) was used. The plate was capped with a septum (Life Technologies) and then heated for 3 minutes at 96 ^{o}C in order to denature the DNA. Formamide also helps in keeping DNA single-stranded. The plate was left at room temperature for 5 minutes before loading to the genetic analyser.

Regarding the instrument, the instructions regarding cleaning, installing the capillary array, performing a spatial calibration and adding polymer were followed according to the manufacturer (Life Technologies, 2014). The dye set G5 (Life Technologies) was selected which contains five dyes that emit fluorescence at different wavelengths (blue 6-FAM™, green VIC®, yellow NED™, red PET®, orange LIZ®). Standard injection and running conditions were used as shown in Table 2-4. DNA profiles were analysed using the GeneMapper ID v.3.2 software (Life Technologies).

Table 2-4. Injection and running conditions in 'FragmentAnalysis36_POP7'

Parameters	Value	Range
Oven temperature (°C)	60	18-65
Polymer filling volume (steps)	6,500	6,500-38,000
Current stability (uAmps)	5	0-2,000
Pre-run voltage (kVolts)	15	0-15
Pre-run time (sec)	180	1-1,000
Injection voltage (kVolts)	1.2	1-15
Injection time (sec)	23	1-600
Voltage number of steps (nk)	20	1-100
Voltage step interval (sec)	15	1-60
Data delay time (sec)	60	1-3,600
Run voltage (kVolts)	15	0-15
Run time (sec)	1,200	300-14,000

2.2.6 DNA sequencing

Once the PCR products were checked through agarose gel electrophoresis to ensure successful amplification with no artefacts, they were ready for sequencing. After amplification, unmethylated cytosines have been transformed to thymines, while the methylated ones remained unchanged. This C/T variation can be treated as equivalent to SNP variation and be analysed with methods that are traditionally used in SNP analysis. Firstly, the main selected method was Pyrosequencing® technology (QIAGEN) as it is very sensitive and allows for successful sequencing of short sequences (~100bp) of the genome and secondly, next generation sequencing using Illumina's MiSeq® platform was employed for multiplex sequencing.

2.2.6.1 Pyrosequencing®

Pyrosequencing® is a real-time DNA sequencing technology based on a luciferase reaction cascade (Ronaghi, 2001). Firstly, the biotin-labelled PCR product, which serves as the Pyrosequencing® template needs to be isolated and denatured in order for the sequencing primer to bind. The solution is then incubated with the enzymes (DNA polymerase, ATP sulfurylase, luciferase and apyrase) and the substrates (adenosine 5' phosphosulfate (APS) and luciferin). Once the deoxynucleotidetriphospate (dNTP) complementary to the base in the template is

dispensed, DNA polymerase facilitates its incorporation into the DNA strand. Every addition is followed by the release of pyrophosphate (PPi), which is proportional to the amount of incorporated nucleotides [Figure 2-5a]. Subsequently, the ATP sulfurylase is able to convert PPi to ATP in the presence of adenosine 5' phosphosulfate, which then initiates the conversion of luciferin to oxyluciferin by luciferase. This conversion generates visible light, which is proportional to the amount of ATP present. The light can be detected by a CCD camera and translated as a peak in the pyrogram™ [Figure 2-5b]. As a result, the height of each peak is directly proportional to the number of nucleotides incorporated, which is proportional to the single-stranded DNA strands in the solution. Lastly, the nucleotide-degrading enzyme, apyrase, continuously degrades ATP and any unincorporated dNTPs before the next dNTP is added. As the sequencing process continues, the complementary DNA strand is created, whose nucleotide sequence can be determined from the pyrogram™ peaks [Figure 2-5c].

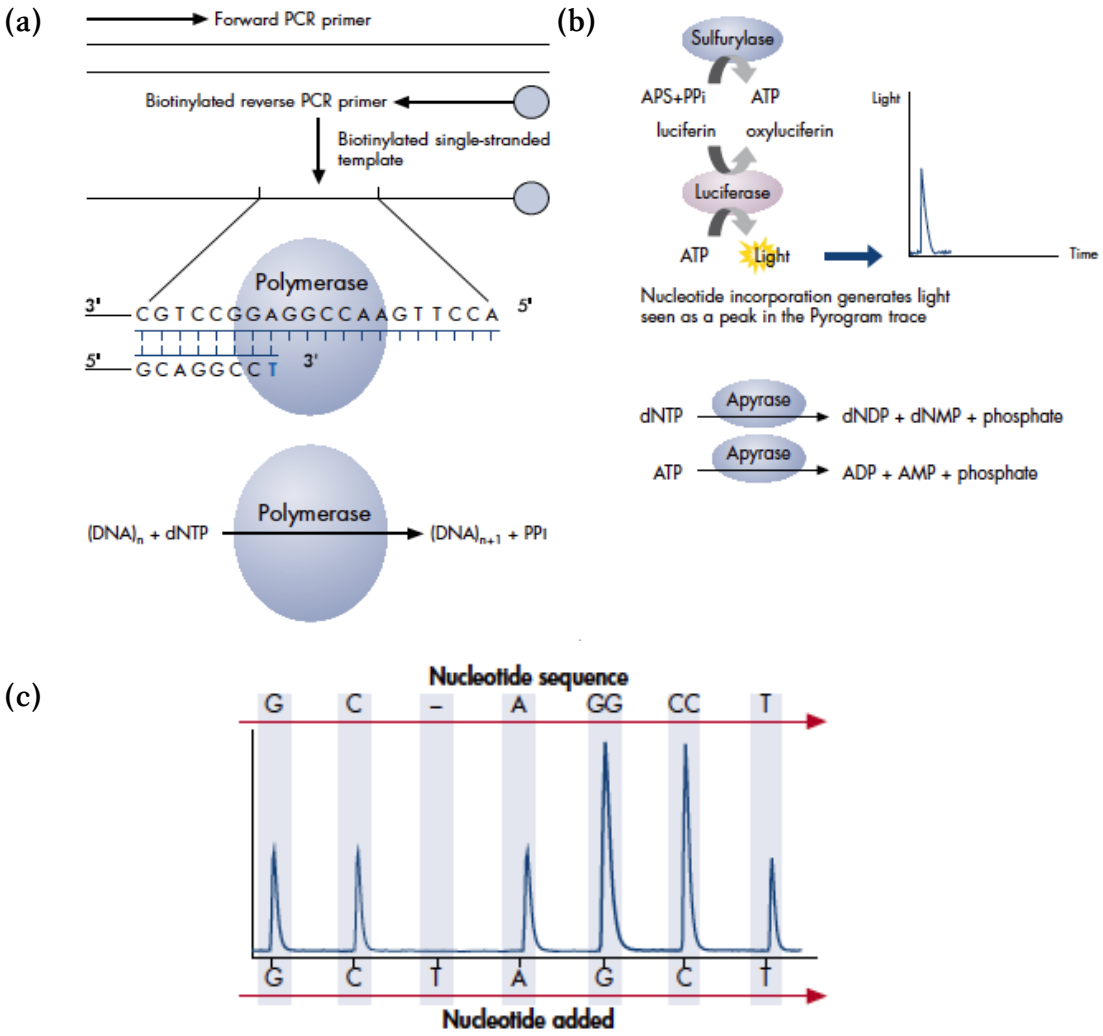


Figure 2-5. The Pyrosequencing® cascade reaction (Qiagen, 2010a)

2.2.6.1.1 Assay design

Pyrosequencing® has the ability to sequence up to 100 bp; however, careful assay design is necessary to minimise artefacts due to events like biotinylated primer interference, non-specific sequencing primer binding or hairpin formation of the single-stranded PCR amplified template. DNA methylation assays were designed in-house including the sequence to analyse and nucleoside dispensation order. All assays were up to 50 bp long as longer reactions were found to be challenging (lower peak signal over time). The CpG of interest and any other adjacent CpG sites were analysed using a sequencing primer that binds up to 5 nucleotides before the first site. Occasionally (especially just before the CpG site of interest), a nucleotide that is not expected to be incorporated was dispensed (also known as dead injection) to ensure that no background interactions take place. Also, at least one bisulphite-treatment control (non-CpG cytosine) was included in the dispensation order; however, a higher number of these controls are preferred to ensure complete conversion of DNA. Figure 2-6 illustrates an example pyrogram™.

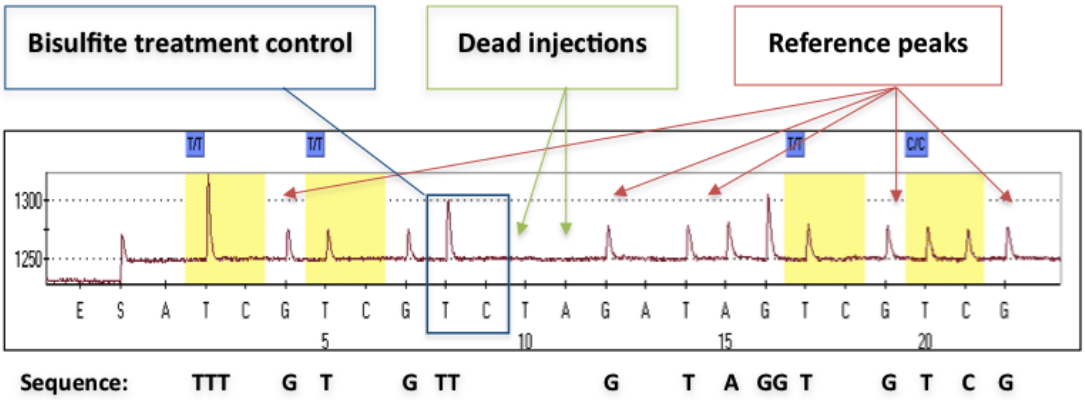


Figure 2-6. Example pyrogram™

The picture illustrates an example pyrogram™ of HBA1 CpG assay that analyses the following bisulphite converted sequence: TTYGYGTTGTAGGYGTYG including four CpG sites in total (marked with yellow). The peak height corresponds to the number of nucleotides that are being incorporated at each injection (for example, if the next bases in the sequence are three Ts – quite common in bisulphite-converted sequences – the peak height will be three times as high as the one produced by only one T). Therefore, the peak heights of the reference peaks (indicated with the red arrows) can reveal the sequence being analysed. Green arrows indicate dead injections – nucleotides being dispensed but not incorporated. Also, the blue box illustrates the bisulphite-treatment control which is a non-CpG cytosine that is expected to be converted during bisulphite treatment. Therefore, a peak for the 'C' dispensation could indicate incomplete conversion (<5% is accepted by the software's settings). Lastly, for each CpG site, both thymine and cytosine are released and peak heights are compared allowing for absolute quantification of methylation.

2.2.6.1.2 Performance assessment

For every new assay that was designed, a few carefully selected controls were run in order to check for any undesirable interaction between the Pyrosequencing® components. These controls were extra Pyrosequencing® reactions that were added to the first run and consisted of annealing buffer and the following: (a) the biotinylated primer (reverse) only, (b) the template only (PCR product), (c) the sequencing primer only and lastly, (d) the biotinylated primer together with the sequencing primer to assess for any primer-primer interactions. If there was background signal, the Pyrosequencing® assay was optimised by either adding less biotinylated primer or adding more PCR product. In cases where optimising the above parameters still resulted in unwanted reactions, the assay was redesigned.

2.2.6.1.3 Pyrosequencing® reaction preparation

In order for the sequencing primer to bind onto the template DNA, the latter has to be single-stranded. This can be achieved via biotin-streptavidin selective binding; 10 µl of PCR products were mixed with 3 µl of Streptavidin Sepharose High Performance Beads (GE Healthcare) and 37 µl of PyroMark Binding Buffer (QIAGEN) as well as 30 µl of distilled water for a total volume of 80 µl. The solutions were then vortexed at 1,000 rpm for 30 minutes using a clear non-skirted 96-well plate (Starlab) to allow for efficient binding of the PCR products to the beads. Afterwards, the beads were isolated and captured utilising a Vacuum Prep Workstation (QIAGEN). This is a simple, specifically-designed preparation tool and consists of a hand-held vacuum prep tool and a vacuum prep worktable with five solution trays connected to a vacuum source.

The principle underlying this bench-top workflow is that the suction caused by the applied vacuum picks up the immobilised PCR products while the rest of the solution is aspirated. While the beads are kept attached to the end of the filter probes of the hand-held tool, they can be cleaned up and denatured in a few steps. The first tray was filled up with 70% ethanol and the probes were washed for 5 seconds in order to remove any undesired leftovers. The second wash was performed using 0.2 M sodium hydroxide (NaOH) for 5 seconds to denature the PCR products and filter out the non-biotinylated amplicon strands. Lastly, the third step involved rinsing the probes with PyroMark Wash buffer for 10 seconds. The single-stranded templates were now ready

so they were transferred to a previously prepared Pyrosequencing® plate (QIAGEN) containing 11.5 µl of PyroMark Annealing buffer (QIAGEN) and 0.5 µl of the appropriate 10 µM sequencing primer. This was achieved by placing the device above the plate, turning the vacuum off and gently shaking the probes until all beads were released into the relevant wells. To ensure complete denaturation of the DNA templates, the plate was also heated at 80 °C for 3 minutes. It was then left at room temperature for 5 minutes in order for the sequencing primer to bind; the plate was then ready to be analysed using a PyroMark MD Pyrosequencer (QIAGEN). All four nucleotides together with the enzymes and substrates (PyroMark Gold Q96 reagents, QIAGEN) were placed into the instrument before analysis following the manufacturer's instructions (Qiagen, 2009c).

2.2.6.2 Next generation sequencing using Illumina MiSeq® platform

Although the potential of Sanger sequencing and Pyrosequencing® has been widely exploited by many laboratories worldwide, it has yet always been hampered by certain limitations in throughput, scalability, speed and resolution. To overcome these barriers and explore its capabilities, a recently developed and fundamentally new technology known as next generation sequencing (NGS) was employed (Illumina, 2013a). In principle, the NGS technology utilises a similar concept, where small DNA fragments are being identified from signals released while they are being re-synthesised using a DNA template. However, scientists are able to extend the sequencing process across millions of reactions in a massively parallel way allowing for rapid sequencing of DNA fragments spanning the entire genome.

For this study, a targeted sequencing approach was chosen, where a subset of selected genomic regions was sequenced in order to further validate reported DNA methylation sites. The ability to pool together multiple samples and obtain high sequence coverage in one single run allowed for simultaneous analysis of up to hundreds of genomic loci. Although various genome-wide DNA methylation studies using Illumina's Infinium HumanMethylation 27K/450K BeadChip arrays have been published, there is only one study that explored the potential of Illumina MiSeq® next generation sequencing system for locus-specific DNA methylation analysis (Masser *et al.*, 2013). Authors reported that they could rapidly and accurately quantify absolute CpG methylation levels from as low as 1 ng of starting DNA material.

2.2.6.2.1 Assay design

For this project, an adjusted version of a protocol that is still at the development stage (not commercially available) was employed (TruSeq Forensic Amplicon, Illumina). The proposed protocol allows for simultaneous sequencing by synthesis of up to 24 DNA libraries in a single reaction. The input DNA is in the form of PCR amplicons (70-300 bp long) that have been previously amplified using ~1 ng of genomic DNA and highly specific, unlabelled PCR primers. The aim is to attach adapter sequences to these PCR amplicons for subsequent cluster generation and sequencing. An overview of sample preparation is illustrated in Figure 2-7. Even though this protocol has been designed to mainly investigate genetic variations (SNPs) that are of forensic interest, in this study it was used to quantify age-associated CpG sites (C/T variation). By integrating six-base sample indexes into the protocol, both loci and samples can be multiplexed. Standard assay design parameters such as the addition of built-in bisulphite conversion controls were followed as described earlier in the chapter.

2.2.6.2.2 PCR input and quantification

It is important to quantify the input PCR product and assess the DNA quality before sample preparation as too much or too little input could result in library preparation failure. 1-100 ng amplified DNA is generally recommended using fluoremetric-based (and not UV spectrophotometric-based) quantification methods like Qubit™ or PicoGreen. Quantification methods that rely on intercalating fluorescent dyes measure only double-stranded DNA (PCR products) and can be less subject to the presence of common contaminants present in the solution such as salts, free nucleotides or proteins. PCR products used in this protocol were prepared as described in section 2.2.4.4 (ZymoTaq™ Premix, ZymoResearch) by amplifying 1 ng of bisulphite-treated DNA. The different PCR products were pooled together before analysis.

<u>Qubit™ dsDNA HS Assay (Life Technologies)</u>

The Qubit™ dsDNA High Sensitivity assay kit (Invitrogen, 2010) allows easy and accurate DNA quantification at room temperature by employing an assay reagent and two pre-diluted DNA standards. DNA concentrations are read by a Qubit™ 2.0 Fluorometer using only thin-wall, clear 0.5 ml PCR tubes.

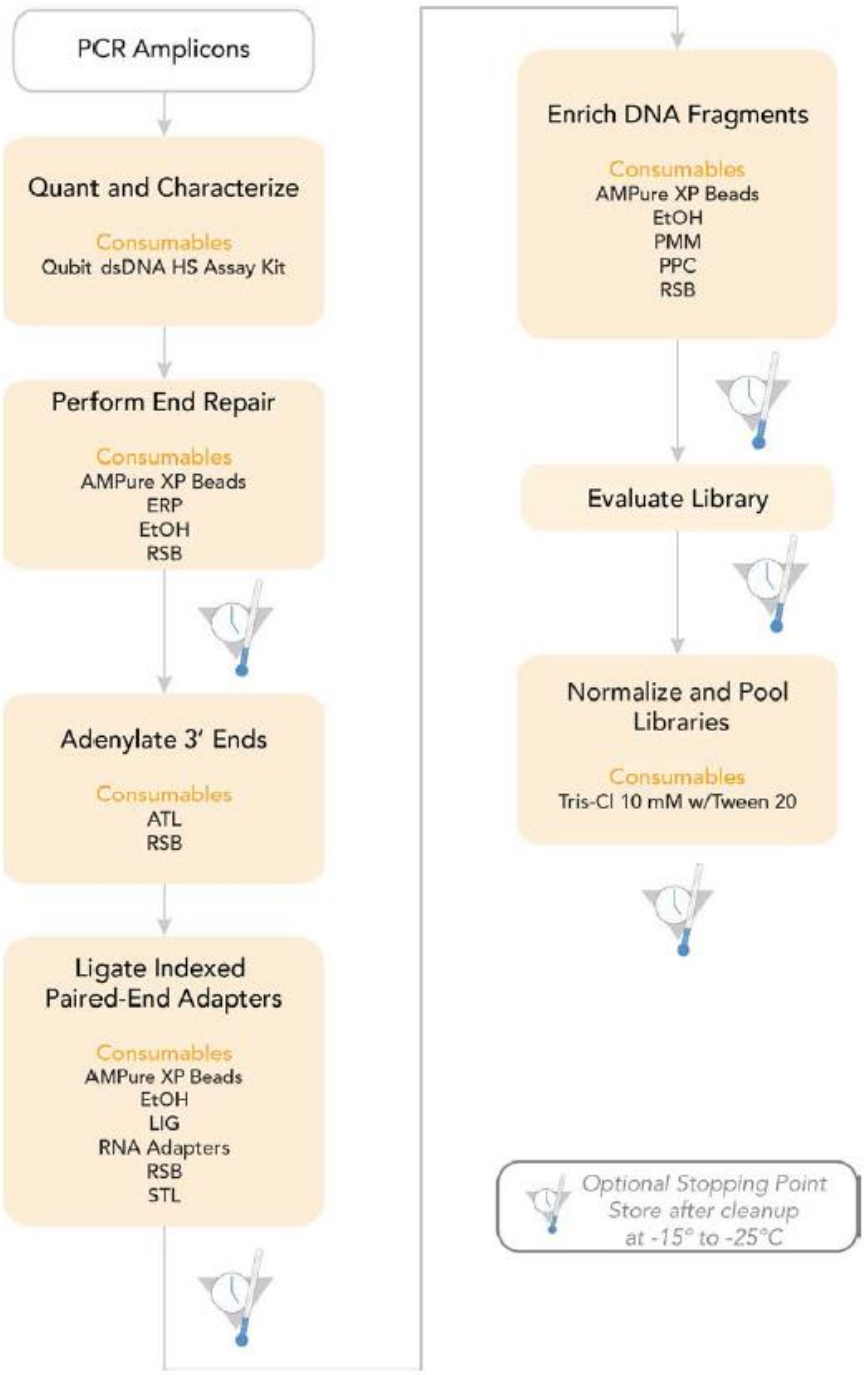


Figure 2-7. TruSeq Forensic Amplicon sample preparation workflow

Firstly, the Qubit™ working solution was prepared by diluting the Qubit™ dsDNA HS reagent 1:200 in Qubit™ dsDNA HS buffer. The final volume of each tube was 200 µl by mixing 190 µl of Qubit™ working solution and 10 µl of each Qubit™ standard or 199 µl of Qubit™ working solution and 1 µl of PCR product. All tubes were mixed by vortexing for 2-3 seconds and left at room temperature for at least 2 minutes. A standard calibration curve was created using the two standards and all measurements for the samples were performed in duplicate. The Qubit™2.0 Fluorometer gives a

concentration value in ng/ml considering the dilution of the assay tube by using the following formula:

Concentration = QF value * (200/x), where

QF value is the value given by the Qubit™ 2.0 Fluorometer and x is the number of microliters of sample added to the assay tube (1 µl). PCR products were normalised to a final volume of 50 µl at 20-2,000 pg/µl (optimum amount for library preparation) and solution were transferred to the wells of a new 96-well 0.3 ml PCR plate.

2.2.6.2.3 Repair of 5' ends

This step converts the 5' overhangs caused by incomplete polymerisation during PCR into blunt ends and phosphorylates 5' ends using an End-Repair mix. For this step, reagents from the TruSeq ChIP Sample Prep kit (Illumina) were used. Briefly, 10 µl of Resuspension buffer (Illumina) were added to each sample containing 50 µl PCR DNA (1-100 ng). 40 µl of End-Repair mix (Illumina) were then added to each well and pipetting of the entire volume up and down ten times was performed to mix thoroughly. The PCR plate was sealed and incubated on a pre-heated Veriti® thermal cycler (Life Technologies) at 30 °C for 30 minutes. Following incubation, 180 µl of well-dispersed AMPure XP Beads (Beckman Coulter Genomics) were added to each well containing 100 µl of End-Repair mix. The solutions were thoroughly mixed by pipetting and incubated at room temperature for 5 minutes.

In order to separate the magnetic beads, the PCR plate was placed on a magnetic stand at room temperature for 15 minutes until the liquid was clear. Then, the supernatant was discarded and some wash steps were performed. While the PCR plate is still on the magnetic stand, 200 µl of freshly-prepared 80% absolute ethanol (Sigma-Aldrich) were added to each well without disturbing the beads. The solutions were left at room temperature for 30 seconds before the supernatant was removed. The wash step was repeated once more for a total of two 80% absolute ethanol washes. It is important that the PCR plate is completely dry before it is taken off the magnetic stand, so it was left for at least 10 minutes at room temperature. Lastly, to resuspend the dried pellets, 17.5 µl of Resuspension buffer was added and the solutions were homogenised by pipetting up and down ten times. Following a 2 minute-incubation at room temperature, the PCR plate was placed back on the magnetic stand until the liquid was clear again. A

total of 15 µl of the clear supernatant from each well was transferred to a new 96-well 0.3 ml PCR plate. The plate could then be stored at -20 °C for up to 7 days before proceeding to the next stage.

2.2.6.2.4 Adenylation of 3' ends

During this process, a single adenine (-A-) nucleotide is added to the 3' ends of the generated blunt fragments in order to prevent them from ligating to each other during the adapter ligation reaction. Also, a corresponding single thymine (-T-) nucleotide on the 3' end of the adapter provides a complementary overhang for ligating the adapter to the fragment. For this step, reagents from the TruSeq ChIP Sample Prep kit (Illumina) were used. In brief, 2.5 µl of Resuspension buffer (Illumina) and 12.5 µl of thawed A-Tailing mix (Illumina) were added to each well of the PCR plate and mixed by pipetting. The plate was sealed and placed on a Veriti® thermal cycler (Life Technologies) for a total of two incubations; 37 °C for 30 minutes followed by 70 °C for 50 minutes. It is essential that the next step is immediately followed.

2.2.6.2.5 Ligation with multiple indexing adapters

This step ensures that multiple indexing adapters are ligated to the ends of the DNA fragments, preparing them for hybridisation on a flow cell. For this step, reagents from the TruSeq ChIP Sample Prep kit (Illumina) were used; it should be noted that there are two sets of RNA Adapter Indexes available shown in Table 5. Before starting the procedure, it is important to completely thaw and centrifuge the RNA adapters.

Table 2-5. Set A and B Indexed adapter sequences (Illumina)

RNA adapter Indexes A	Sequence	RNA adapter Indexes B	Sequence
A002	CGATGT	**A001**	ATCACG
A004	TGACCA	**A003**	TTAGGC
A005	ACAGTG	**A008**	ACTTGA
A006	GCCAAT	**A009**	GATCAG
A007	CAGATC	**A010**	TAGCTT
A012	CTTGTA	**A011**	GGCTAC
A013	AGTCAA	**A020**	GTGGCC
A014	AGTTCC	**A021**	GTTTCG
A015	ATGTCA	**A022**	CGTACG
A016	CCGTCC	**A023**	GAGTGG
A018	GTCCGC	**A025**	ACTGAT
A019	GTGAAA	**A027**	ATTCCT

Briefly, 2.5 µl of Resuspension buffer (Illumina) together with 2.5 µl of Ligation mix (Illumina) were added to each well of the PCR plate. It is important that the ligation mix is returned to -20 °C immediately after use. Afterwards, 2.5 µl of each RNA Adapter Index was added to the corresponding well and mixed with pipetting. The plate was sealed and centrifuged at 280 g for 1 minute to collect any drops on the walls of the tubes. The solutions were then incubated at 30 °C for 10 minutes and the ligation process was inactivated by adding 5 µl of Stop Ligation buffer.

The second part of this step includes the clean-up and resuspension using the magnetic beads. 59.5 µl of well-dispersed AMPure XP Beads (Beckman Coulter Genomics) were added to each well of the PCR plate, which was then left at room temperature for at least 5 minutes. In order to separate the magnetic beads, the PCR plate was placed on a magnetic stand at room temperature for 5 minutes until the liquid was clear. Then, the supernatant was discarded and some wash steps were performed. While the PCR plate was still on the magnetic stand, 200 µl of freshly-prepared 80% absolute ethanol (Sigma-Aldrich) was added to each well without disturbing the beads. The solutions were left at room temperature for 30 seconds before the supernatant was removed. The wash step was repeated once more for a total of two 80% absolute ethanol washes. It is important that the PCR plate is completely dry before it is taken off the magnetic stand, so it was left to air-dry.

To resuspend the dried pellets, 52.5 µl of Resuspension buffer was added and the solutions were homogenised by pipetting up and down ten times. Following a two-minute incubation at room temperature, the PCR plate was placed back on the magnetic stand until the liquid was clear again. A total of 50 µl of the clear supernatant from each well was transferred to a new 96-well 0.3 ml PCR plate. The above procedure was repeated once more by using 70 µl of well-dispersed AMPure XP Beads and following the same steps. The dried pellets were resuspended by adding 22.5 µl of Resuspension buffer and finally, a total of 20 µl of the clear supernatant from each well was transferred to a new 96-well 0.3 ml PCR plate. The plate could then be stored at -20 °C for up to 7 days before proceeding to the next stage.

2.2.6.2.6 Enrichment of DNA fragments

The last step of the library preparation includes the enrichment via PCR of the DNA fragments with adapter molecules on both ends. Fragments with only one or no adapters are by-products of inefficiencies during the ligation reaction. This PCR is performed with the PCR Primer Cocktail that anneals to the ends of the adapters. To retain library representation, the number of PCR cycles was kept to the minimum. For this step, reagents from the TruSeq ChIP Sample Prep kit (Illumina) were also used.

5 µl of thawed PCR Primer Cocktail (Illumina) was added to each well, followed by the addition of 25 µl of PCR Master mix (Illumina). Solutions were mixed thoroughly by pipetting up and down ten times and the plate was then sealed and placed on a thermal cycler. The following program was used: 98 °C for 30 seconds followed by 18 cycles of 98 °C for 10 seconds, 60 °C for 30 seconds and 72 °C for 30 seconds and a final extension step at 72 °C for 5 minutes. Following incubation, 50 µl of well-dispersed AMPure XP Beads (Beckman Coulter Genomics) was added to each well containing 100 µl of End-Repair mix. The solutions were thoroughly mixed by pipetting and incubated at room temperature for 5 minutes.

The PCR plate was then placed on a magnetic stand at room temperature for 15 minutes until the liquid was clear. Then, the supernatant was discarded and some wash steps were performed. While the PCR plate was still on the magnetic stand, 200 µl of freshly-prepared 80% absolute ethanol (Sigma-Aldrich) were added to each well without disturbing the beads. The solutions were left at room temperature for 30 seconds before the supernatant was removed. The wash step was repeated once more for a total of two 80% absolute ethanol washes. It is important that the PCR plate is completely dry before it is taken off the magnetic stand, so it was left to air-dry.

Solutions were resuspended by adding 32.5 µl of Resuspension buffer were added and the solutions were homogenised by pipetting up and down ten times. Following two-minute incubation at room temperature, the PCR plate was placed back on the magnetic stand until the liquid was clear again. A total of 30 µl of the clear supernatant from each well was transferred to a new 96-well 0.3 ml PCR plate. The plate could then be stored at -20 °C for up to 7 days before proceeding to the next stage.

2.2.6.2.7 Evaluation of generated libraries

In order to achieve the highest quality data on the MiSeq® sequencing platform, it is essential that the optimal cluster density across all lanes of the flow cell is achieved. Therefore, it is a requirement that the DNA library templates are accurately quantified. There are two options available, one including a qPCR assay (KAPA Biosystems) and the other one using a Bioanalyzer (Agilent Technologies) and the High Sensitivity DNA chip. For this study, libraries were evaluated using the KAPA Library quantification kit for Illumina sequencing platforms as qPCR specifically quantifies only PCR-competent DNA molecules and is highly sensitive allowing for accurate quantification of low concentration libraries (KAPABiosystems, 2014).

Briefly, 1 ml of Illumina Primer premix (10X) was mixed with 5 ml of KAPA SYBR® FAST qPCR Master mix to prepare the qPCR/Primer mix. dsDNA libraries were diluted 1:2000 in 10 mM Tris-HCl, pH 8.0 + 0.05% Tween 20 (Sigma-Aldrich) (total volume of 2,000 µl). For each 10 µl reaction, 6 µl of the qPCR/Primer mix were mixed with 4µl of the diluted DNA library or DNA standards (supplied in the kit, concentration range 20-0.0002 pM). All reactions were performed in duplicate. The qPCR protocol included an initial activation/denaturation step at 95 °C for 5 minutes followed by 35 cycles of 95 °C for 30 seconds and 60 °C for 30 seconds.

The obtained concentrations were adjusted using the average fragment length and the following formula:

Final concentration (pM) = Average concentration (pM) * (452/Average fragment length) * 2000

2.2.6.2.8 Normalisation and cluster generation

This step describes how DNA templates are prepared for cluster generation. Indexed DNA libraries were normalised to 10 nM using Tris-HCl 10 mM, pH 8.5 with 0.1% Tween 20. 10 µl of the sample libraries from each well of the PCR plate were pooled for a total volume of 240 µl. The entire normalised library was gently mixed by pipetting up and down ten times.

2.2.6.2.9 Library dilution for sequencing on the MiSeq®

This step explains how to denature and dilute libraries after sample preparation to prepare them for sequencing on the MiSeq® platform including the preparation of the PhiX control. The latter is important when analysing low-diversity libraries – libraries where a significant number of reads have the same sequence. Using the PhiX control prevents potential shifting of the base composition due to the reads being no longer random. Firstly, 1 ml of 0.2 N NaOH was freshly prepared before each run by mixing 800 µl of laboratory-grade water and 200 µl of stock 1.0 N NaOH. A fresh solution is required to ensure efficient denaturation of the libraries. Next, libraries were further diluted down to 4 nM; by mixing 5 µl of the diluted libraries (4 nM) with 5 µl of 0.2 N NaOH. The solution was mixed by vortexing and left at room temperature for at least 5 minutes to obtain single-stranded DNA. 990 µl of pre-chilled Hybridisation buffer (HT1) (Illumina) was then mixed with 10 µl of the denatured DNA, which results in 20 pM denatured libraries in 1 mM NaOH. However, libraries should be further diluted to give 600 µl of the desired input concentration as follows:

Final Concentration	**6 pM**	**8 pM**	**10 pM**	**12 pM**	**15 pM**	**20 pM**
20 pM denatured DNA (µl)	180	240	300	360	450	600
HT1 buffer (µl)	420	360	300	240	150	0

The 10 nM PhiX control library was prepared to 20 pM. Firstly, 2 µl of the provided 10 nM Phix library were mixed with 3 µl of 10 mM Tris-HCl, pH 8.5 with 0.1% Tween 20. Then, 5 µl of the 4 nM PhiX library were added to 5 µl of 0.2 N NaOH and incubated for at least 5 minutes at room temperature to ensure complete denaturation. Lastly, to obtain a 20 pM PhiX library the above 10 µl denatured library was mixed with 990 µl of pre-chilled HT1 Hybridisation buffer. Illumina recommends a low-concentration PhiX control spike-in at a minimum 5% for low-diversity libraries. Therefore, 60 µl of the denatured and diluted PhiX control was mixed with 540 µl of the sample library for a final volume of 600 µl. The preparation of the pre-filled, reagent cartridge and flow cell (Illumina) and the set-up of the instrument were followed according to the manufacturer's instructions (Illumina, 2013b). It should be noted that the instrument was set up to run a paired-end read of 150 bp of DNA sequence from both ends of the library products. Data was entered in the sample sheet to link each sample with the corresponding ligated RNA adapter. Auto analysis was set up as a FASTQ-only method.

2.3 RNA analysis

RNA was extracted and analysed for the purpose of mRNA profiling for body fluid/tissue identification.

2.3.1 DNA/RNA co-extraction

Since DNA is essential for the identification of a body fluid stain's donor, the method to be selected for RNA extraction should allow for simultaneous isolation of both nucleic acids. Similarly with the sole DNA extraction, body fluid samples were either in a liquid form, deposited on a swab or on a surface (fabric, glass). The whole of these swabs or stains was used to co-extract DNA and RNA from cellular material. The method used in this study was the AllPrep DNA/RNA mini kit (QIAGEN).

2.3.1.1 AllPrep DNA/RNA mini (QIAGEN)

The AllPrep DNA/RNA mini kit allows for simultaneous purification of genomic DNA and total RNA from a single cell or tissue sample. The method's principle is based on the fact that lysate is first passed through an AllPrep DNA spin column to selectively isolate DNA and then through an RNeasy spin column to selectively bind RNA [Figure 2-8]. Depending on the homogenisation conditions, purified DNA usually has an average length of 15-30 kb, while the general procedure allows for the isolation of RNA molecules longer than 200 nucleotides. This results in an enrichment of mRNA, since most <200 bases RNAs (such as rRNAs and tRNAs) are carefully removed. The protocol can be adjusted depending on the type of starting material and more information can be found in the relevant handbook (Qiagen, 2005).

In this study an adjusted version of the protocol 'Simultaneous purification of genomic DNA and total RNA from animal cells' was employed. Before any extraction was performed, all bench surfaces, pipettes and consumables were sprayed with RNaseZap® RNase decontamination solution (Life Technologies) to prevent possible RNA degradation by contaminating, free RNases. Additionally, 10 µl of ß-mercaptoethanol (ß-ME) (Sigma-Aldrich) was added to 1 ml of buffer RLT Plus (QIAGEN) in order to ensure sufficient cell lysis. According to the manufacturer, carrier RNA (1 µg/µl) (QIAGEN) was used to enhance DNA binding onto the membrane. The carrier RNA working solution was made up as follows; firstly, 2.5 µl

of carrier RNA was mixed with 17 µl of buffer RLT Plus and then, the solution was further diluted ten times.

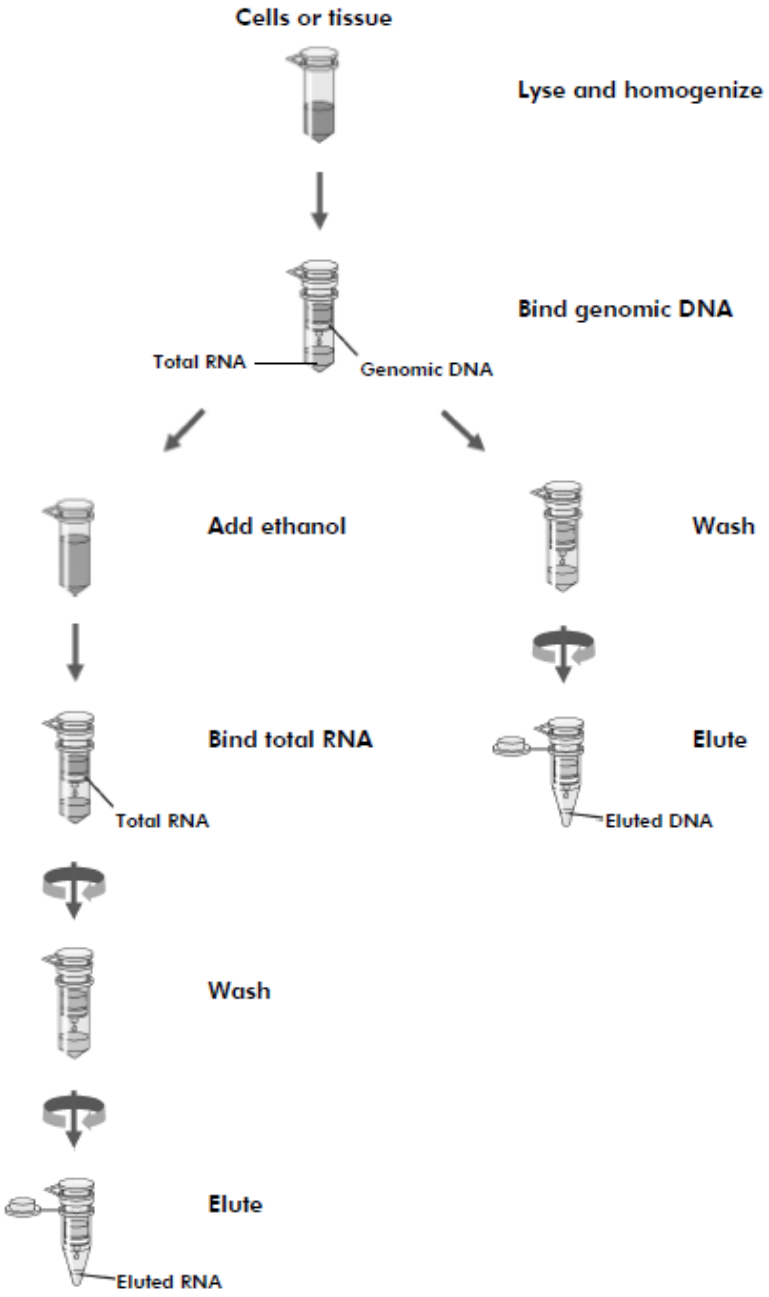


Figure 2-8. Overview of the AllPrep DNA/RNA Procedure (Qiagen, 2005)

Briefly, the stains or swabs were cut using sterile scissors or a scalpel; in case of a 'touch' sample or where the stain was not visible, a swab was moistured with 50 µl of RNase-free water and used to swab the 'suspected' area. Cotton swabs or stains were transferred to a clean 1.5 ml tube and 345 µl of RLT buffer Plus with ß-ME as well as 5 µl of carrier RNA working solution were added. The samples were mixed by vortexing for at least 30 seconds followed by incubation at 56 °C for up to 3 hours.

Once incubation was complete, suitable spin baskets (DNA IQ™ spin basket, Promega) were applied to dry out any material on the swab/cotton material by centrifuging at 13,000 rpm for 5 minutes. Lysates were firstly transferred to the DNA spin columns (violet) and centrifuged at 11,000 rpm for 30 seconds. As mentioned in Figure 2-8, DNA binds to the membrane of the columns, while RNA passes through the flow-through. Regarding RNA isolation, the flow-through samples were mixed with 350 µl of freshly-made 70% ethanol and the solutions were mixed by pipetting up and down ten times. The solutions were then transferred to the RNA spin columns (pink) and centrifuged at 11,000 rpm for 30 seconds. The flow-through was discarded since RNA was bound onto the column membrane. 700 µl of RW1 buffer (QIAGEN) were added to the columns, which were then centrifuged at 11,000 rpm for 15 seconds. The spin columns were placed in new collection tubes and 500 µl of RPE buffer (QIAGEN) was added followed by centrifugation at 11,000 rpm for 30 seconds. This step was repeated once more. Spin columns were centrifuged at maximum speed for 1 minute to ensure that membranes had completely dried out. RNA was eluted by applying 30 µl of RNase-free water to the centre of the membrane, incubating at room temperature for 5 minutes and centrifuging at 11,000 rpm for 1 minute. In suspected low-quantity samples, the solution was reloaded back to the column and the elution step was repeated to maximise RNA yield.

Regarding DNA isolation, 500 µl of AW1 wash buffer (QIAGEN) was added to the DNA spin columns, which were then centrifuged at 11,000 rpm for 15 seconds. Another 500 µl of AW2 wash buffer was added and solutions were centrifuged at maximum speed (14,000 rpm) for 2 minutes to completely wash and dry the membranes. DNA was eluted by applying 40-80 µl to the centre of the membrane, incubating at room temperature for 1 minute and centrifuging at 11,000 rpm for 1 minute. Both RNA and DNA samples were stored at -20 ^{o}C until analysis.

2.3.2 DNase treatment

Even though the above protocol promises DNA and RNA separation, extracted RNA samples might contain contaminating DNA traces, which are likely to interfere with downstream analysis. Therefore, it is essential to remove any DNA present in the RNA samples by using DNase. Preferably, this step should follow immediately after RNA extraction.

2.3.2.1 TURBO™ DNA-free kit (Ambion)

For this study, the TURBO™ DNA-free kit was used (Ambion, 2012). Briefly, 30 µl of RNA solutions were mixed with 3 µl (0.1 of RNA volume) of 10X TURBO™ DNase buffer (Ambion) and 1 µl of TURBO™ DNase (Ambion) followed by incubation at 37 ^{o}C for 30 minutes. 3.4 µl (0.1 volume) of DNase Inactivation reagent (Ambion) were then added to inactivate the enzyme and solutions were incubated at room temperature for 5 minutes while mixed a few times by vortexing. Following incubation, RNA solutions were centrifuged at 10,000 rpm for 90 seconds; the inactivation reagent forms a visible, white pellet. Approximately 32 µl were then transferred to a new 1.5 ml tube. DNA-free RNA solutions were stored at -20 ^{o}C until needed.

2.3.3 Synthesis of complementary DNA (cDNA)

It is important that mRNA is converted to cDNA so that the latter can be analysed by means of reverse transcription PCR (RT-PCR). Complementary DNA (cDNA) is the DNA synthesised using the mRNA as a template in a reaction catalysed by an enzyme known as reverse transcriptase. In general, cDNA can be synthesised using either total RNA or poly(A)$^{+}$-selected RNA primed with oligo(dT), random primers, or a gene-specific primer. Random hexamers are the most non-specific priming method and are usually employed when the mRNA is difficult to copy in its entirety (due to low quality/quantity). Therefore, in this study, using random primers, all RNAs in a population could be used as templates for cDNA synthesis. Also, it was suggested that up to 40% of extracted RNA is used for cDNA synthesis.

2.3.3.1 SuperScript® III First-Strand Synthesis system (Invitrogen)

In this project, the SuperScript® III First-Strand Synthesis system for RT-PCR was employed according to the manufacturer's instructions (Invitrogen, 2013). Using this protocol, RNA targets from 100 bp to >12 kb can be detected and the amount of starting material can vary from 1 pg to 5 µg of total RNA. All reagents mentioned in this section were included in the kit. Briefly, 12 µl of DNA-free RNA solutions were mixed with 4 µl of RNase-free water, 2 µl of 10mM dNTP mix and 2 µl of random primers (2.5 ng/µl) for a total volume of 20 µl. The solutions were incubated at 65 ^{o}C for 5 minutes and were immediately placed on ice for at least 1 minute. For each

reaction, the following reagents were used: 4 μl of 10X RT buffer, 8 μl of 25 mM $MgCl_2$, 4 μl of 0.1 M DTT and 2 μl of RNaseOUT™ (40 U/μl). All reagents were mixed well by vortexing and added to the 20 μl of RNA/primer mixture for a total volume of 38 μl. To create the RT+ and RT- controls, each reaction was divided into two tubes (19 μl each) and either 1 μl of SuperScript® III RT (200 U/μl) or 1 μl of RNase-free water was added. cDNA synthesis was performed as follows; incubation at 25 °C for 10 minutes followed by an incubation at 50 °C for 50 minutes. Reactions were terminated by heating the samples at 80 °C for 5 minutes. By adding 1 μl of RNase H to each tube, any RNA left could be eliminated. cDNA samples were stored at -20 °C until further analysis.

2.3.4 RT-PCR

Regarding the amplification of complementary DNA, standard PCR parameters can be applied as previously mentioned in section 2.2.4. However, it should be noted that one of the most important assay design parameters in this type of analysis is the primer design, as co-amplification of contaminating genomic DNA should be avoided. Therefore, RT-PCR primers were designed to span an exon-to-exon boundary. In this study, all used primers together with PCR guidelines were sent by the collaborating laboratories. Details on the PCR set up and cycling conditions of each assay tested can be found in Chapter 4.

2.3.5 Post-PCR product purification

Following amplification, RT-PCR products can be purified to eliminate excess primer and dNTPs or any other small molecule present in the solution that could affect downstream analysis. Using post-PCR product purification, amplicons can also be concentrated by decreasing the elution volume. It should be noted that only fragments <40 bp are removed, therefore PCR products should be at least 70 bp long to ensure successful recovery.

2.3.5.1 MinElute PCR purification (QIAGEN)

In this study, the MinElute PCR purification kit was used according to manufacturer's conditions (Qiagen, 2008). The system is specially designed for fast and easy isolation of DNA fragments offering an extremely small elution volume of only 10 μl. The method is based on the selective binding properties of a uniquely designed silica

membrane that allows the purification of 70 bp – 4 kb long DNA fragments. DNA adsorbs to the silica membrane in the presence of high concentrations of salt while contaminants pass through the column. Impurities are eventually washed away and the pure DNA is eluted with Tris buffer or water (~80% recovery).

Briefly, 125 μl of Buffer PB (QIAGEN) were added to 25 μl of PCR product and the solutions were mixed by pipetting. They were then transferred to the MinElute columns and centrifuged at maximum speed (14,000 rpm) for 1 minute. A washing step was performed by adding 750 μl of Buffer PE (QIAGEN) and centrifuging at maximum speed (14,000 rpm) for 1 minute. Columns were dried out by centrifugation for another minute. DNA fragments were eluted by adding 10 μl of buffer EB (QIAGEN) to the centre of the membrane, incubating at room temperature for 1 minute and centrifuging at maximum speed (14,000 rpm) for 1 minute. Purified PCR products were stored at -20 °C.

2.3.6 cDNA fragment analysis

PCR products were separated and analysed using capillary electrophoresis as described in section 2.5.2.2. Similarly with STR amplicons, PCR products were fluorescently labelled with various dyes and could be successfully identified by their expected fragment length. The adjusted CE parameters are illustrated in Table 2-6.

Table 2-6. Injection and running conditions for cDNA fragment analysis

Parameters	Value	Range
Oven temperature (°C)	60	18-65
Polymer filling volume (steps)	6,500	6,500-38,000
Current stability (uAmps)	5	0-2,000
Pre-run voltage (kVolts)	15	0-15
Pre-run time (sec)	180	1-1,000
Injection voltage (kVolts)	3	1-15
Injection time (sec)	10	1-600
Voltage number of steps (nk)	40	1-100
Voltage step interval (sec)	15	1-60
Data delay time (sec)	1	1-3,600
Run voltage (kVolts)	15	0-15
Run time (sec)	1,800	300-14,000

2.4 Data analysis

This section describes the main methods used in this project regarding data analysis.

2.4.1 DNA methylation data

2.4.1.1 Identification of CpG chromosomal locations

Once specific CpG sites had been selected for analysis, the online Ensembl genome browser and in particular the Human (*Homo sapiens*) Feb 2009 (GRCh37/hg19) genome was used to obtain the required genetic information for assay design (http://www.ensembl.org/index.html). Their exact chromosomal locations were confirmed and the surrounding DNA sequences were identified. Information on the genes and their function as well as known SNPs included in the regions of interest was also obtained. DNA sequences and locations were also verified using a second online browser, the UCSC genome browser (http://genome.ucsc.edu/). An area of around 500 bp spanning each CpG of interest was used for bisulphite PCR assay design.

2.4.1.2 Pyrosequencing® data analysis

This section will describe how methylation data obtained by Pyrosequencing® was analysed assuming that pyrograms™ passed the instrument's quality control. This included the detection of desired peaks (no signals in dead injections) as well as detected bisulphite conversion rates of >95%. For each CpG site the degree of methylation was measured as a frequency of C/T signals in the form of peak heights. Pyrosequencing® dispenses both thymine (T) and cytosine (C) for each CpG site; the peak height of thymine corresponds to the unmethylated sequence (C), whereas the peak height of cytosine corresponds to the methylated sequence (mC). In the initial Pyrosequencing® assay development, standard software used for SNP variation analysis, PyroMark Q96 MD software was used. For this, methylation levels had to be calculated manually by the ratio of mC:C at each CpG site using the following formula:

% Methylation = [(peak height C)/(peak height C + peak height T)] * 100

There were cases where the peak for the unmethylated cytosines (T signal) of a CpG site coincided with the peak of other Ts including converted non-CpG cytosines. In these cases, the average peak height was calculated taken into account all peaks

included in the pyrogram™ and subtracted from the obtained T signal as many times as the number of non-CpG sites included. It was understood that manual calculation could result in inaccuracies, thus CpG methylation software, PyroMark CpG SW 1.0 software (QIAGEN) was then employed. This software has been specifically developed for DNA methylation analysis and uses an algorithm for calculating methylation levels. One *versus* multiple peak effect did not impact either the resulting peak height calls or the reported methylation values as the software accounts for such effects.

2.4.1.3 MiSeq® data analysis

This section will describe how methylation data generated by Illumina's next generation sequencing platform was analysed assuming that methylation profiles passed the instrument's quality control. Firstly, auto-analysis by the MiSeq® Reporter Software separated the millions of generated sequences into the constituent samples on the basis of the ligated adaptor tags that were unique to each sample in the library. These collated sequences were packaged in a text-based format (FASTQ files), with each sample associated with two of these FASTQ files, one for sequences generated during the forward read and one for those produced during the reverse read. FASTQ is a modification of a FASTA file to include quality data as well as sequence.

Sequences within these FASTQ files were aligned to the previously created custom genome using a Burrows-Wheeler alignment (BWA) algorithm. This process was implemented in the BWA program (Li & Durbin, 2010) using the maximum entropy method (MEM) algorithm that matched the sequences generated to the respective methylation marker (i.e. a sequence obtained from the PCR product of any specific marker would be most similar to the reference sequence for that marker, and hence the software would align this sequence with that marker). In the case of bisulphite sequencing, in order to create the custom reference genome (DNA sequences of all markers) that the generated sequences would align to, one has to make certain assumptions. These inlcude the assumption that all non-CpG cytosines had been converted to thymines following bisulphite conversion and PCR, and also that all CpG cytosines were methylated, therefore remained cytosines during analysis (Appendix IV). It should be noted that no difference in the reported methylation levels was observed when assuming all CpG sites in the amplicons were unmethylated. In this

way, the millions of sequences contained within the FASTQ file can be associated with their respective marker, giving potentially hundreds of thousands of individual sequences all aligned in parallel to a specific reference sequence/marker. At the conclusion of this alignment process, a sequence alignment/map (SAM) file was produced, which was further modified using SAM tools (Li *et al.*, 2009) to convert it into a BAM file.

The Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010) was subsequently used to interrogate this BAM file by targeting specific positions in these aligned sequences (i.e. at each CpG site) and reporting the number of sequences containing a C and the number of sequences containing a T at this position. In this way it was possible to assess the methylation state at every studied CpG site. The unified genotyper algorithm was employed within GATK to produce this genotype data for each CpG site, and this data was written by the software into a variant call format (vcf) file that could subsequently be manipulated in Excel. The command script for performing this analysis written by Dr David Ballard at KCL can be found in Appendix III.

2.4.1.4 Calculation of bisulphite conversion rates

As mentioned earlier when discussing assay design, each assay included at least one non-CpG cytosine as a bisulphite conversion control. These cytosines do not belong to a CpG site; therefore, they are expected to be converted completely during bisulphite treatment. By investigating their methylation, the bisulphite conversion efficiency could be assessed; absence of C signals for these locations would indicate 100% conversion. Although the Pyrosequencing® software is set up to accept at least 95% conversion rates, in some cases (MiSeq® data) these rates were calculated manually to assess suspected lower rates. Using the procedure followed when quantifying CpG methylation state, the ratio of mC:C was calculated.

2.4.1.5 Methylation correction by Mathematica

For all developed CpG assays, DNA standards of known methylation levels (0%-100%) were used to assess the linearity of methylation quantification obtained using either the Pyrosequencing® or MiSeq® instrument. The observed methylation ratio was then plotted against the expected and the best-fitted regression line was chosen.

As it will be detailed in the Chapters 5, 7 and 8, in many cases the association was not linear. It is known from previous studies that PCR amplification of bisulphite-treated DNA can often result in a selective enrichment of unmethylated alleles (PCR bias) (Moskalev *et al.*, 2011); preferential recovery of methylated forms is also possible, although less common. Differences in PCR efficiency are mainly due to differences in DNA sequence between the unmethylated and methylated templates following bisulphite conversion, which are translated in differences in CG content. Differences in hydrogen bonding between A-T and C-G pairs (two and three hydrogen bonds respectively) could lead into different separation and amplification rates. For example, in case that the methylated allele is preferentially amplified, the resulting graph is curved [Figure 2-9d], affecting mainly the low-methylated DNA templates and reaching a plateau when highly methylated DNA is present (since the competition between unmethylated and methylated templates is in that case lower). The extent of such deviations is difficult to predict and can often still remain after extensive PCR optimisation.

Applying the criterion of minimising the sum of squared error, quadratic or cubic polynomial regression lines were found to fit best and substantially improve the association between the observed and expected methylation ratios. Therefore, the obtained equations of these regression lines were used to 'correct' the biased detected methylation levels in order to increase the accuracy of the proposed methods. To achieve this, a computational software program called MATHEMATICA v4.1.2 (Wolfram Research Europe) was used. Using the appropriate feature, polynomial equations were solved to provide the corrected methylation values. If the corrected methylation resulted in negative values, methylation was considered as zero. An example of this correction for CpG site cg19761273 is illustrated in Figure 2-9.

2.4.2 Fragment data

The resulting electropherograms from fragment analysis applications (either STR or RNA analysis) were obtained and interpreted using GeneMapper ID v3.2 software (Applied Biosystems). The software provides the fragment size of sample peaks by comparing them to the synthetic size standard fragments run along with each sample.

(a) Cubic polynomial equation $y = ax^3 + bx^2 + cx + d$
where y = observed, and x=true methylation value
Obtained equation for cg19761273: $y = 1.2669x^3 - 3.0033x^2 + 2.6948x + 0.0255$

(b)

cg19761273	DNA methylation levels for each DNA standard						
	0	0.05	0.1	0.25	0.5	0.75	1
Expected	0.0048	0.0493	0.0981	0.2465	0.4930	0.7394	0.9811
Observed	0.0074	0.1495	0.3297	0.4836	0.7793	0.8952	0.9721
Parameter d	0.0181	-0.1240	-0.3042	-0.4581	-0.7538	-0.8697	-0.9466
Corrected	-0.00667	0.04859	0.13094	0.218151	0.49823	0.76068	0.97488

(c)

```
In[4]:= Solve[ 1.2669 x^3 - 3.0033 x^2 + 2.6948 x + 0.0181 == 0 , x]
Out[4]= {{x → -0.00666696}, {x → 1.18863 - 0.854455 i}, {x → 1.18863 + 0.854455 i}}
In[5]:= Solve[ 1.2669 x^3 - 3.0033 x^2 + 2.6948 x - 0.1240 == 0 , x]
Out[5]= {{x → 0.0485921}, {x → 1.161 - 0.816292 i}, {x → 1.161 + 0.816292 i}}
In[6]:= Solve[ 1.2669 x^3 - 3.0033 x^2 + 2.6948 x - 0.3042 == 0 , x]
Out[6]= {{x → 0.130936}, {x → 1.11983 - 0.761458 i}, {x → 1.11983 + 0.761458 i}}
        Solve[ 1.2669 x^3 - 3.0033 x^2 + 2.6948 x - 0.4581 == 0 , x]
Out[7]= {{x → 0.218151}, {x → 1.07622 - 0.706595 i}, {x → 1.07622 + 0.706595 i}}[1]
In[8]:= Solve[ 1.2669 x^3 - 3.0033 x^2 + 2.6948 x - 0.7538 == 0 , x]
Out[8]= {{x → 0.498233}, {x → 0.936178 - 0.563722 i}, {x → 0.936178 + 0.563722 i}}
In[9]:= Solve[ 1.2669 x^3 - 3.0033 x^2 + 2.6948 x - 0.8697 == 0 , x]
Out[9]= {{x → 0.760675}, {x → 0.804957 - 0.504484 i}, {x → 0.804957 + 0.504484 i}}
In[10]:= Solve[ 1.2669 x^3 - 3.0033 x^2 + 2.6948 x - 0.9466 == 0 , x]
Out[10]= {{x → 0.697857 - 0.528611 i}, {x → 0.697857 + 0.528611 i}, {x → 0.974876}}
```

(d)

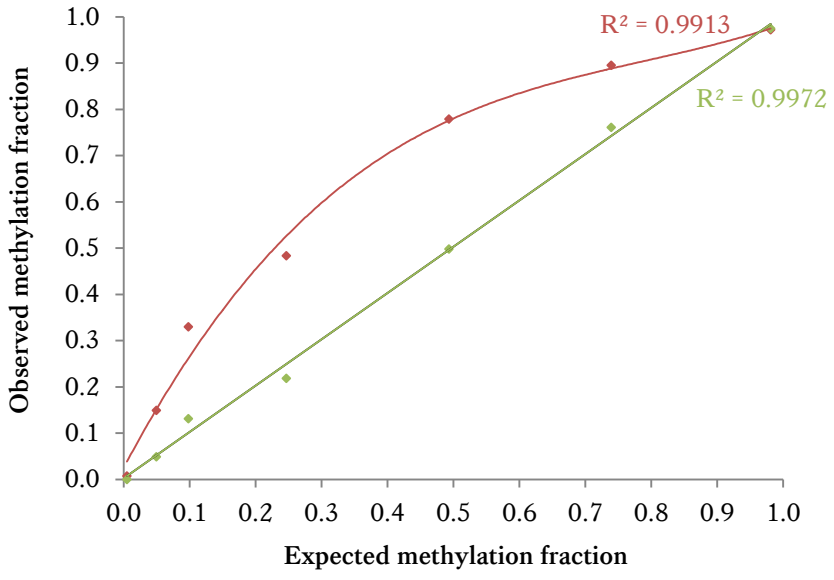


Figure 2-9. Correction of methylation values for cg19761273 using MATHEMATICA
(a) Cubic polynomial equation obtained following analysis of pre-defined DNA methylation standards, (b) Observed, expected and corrected methylation values, (c) solutions of cubic polynomial equations for each DNA methylation standard, (d) observed *vs.* expected methylation graphs obtained before (red, cubic polynomial regression) and after (green, linear regression) methylation correction.

For STR applications, the fragment sizes can be converted into alleles (exact number of STR repeats) by applying the allelic ladder. An allelic ladder is a collection of all possible, common alleles, each of which has a verified repeat number. While being separated by the instrument, each known allele will migrate at a particular rate having a specific size allowing for comparison with the sizes obtained from the samples. A size standard should be added to every injection to correct for any potential fluctuations of the instrument's performance. Sample peaks were assigned a particular allele number as long as the corresponding peak in the allelic ladder was not more than ±0.5 bp different.

Lastly, for RNA applications, the fragment sizes of the PCR products were known and used to successfully identify each mRNA marker. A peak of >50 rfu height would indicate the presence of a particular mRNA. Specifically for the TissueID 20plex, bins were set up in the software corresponding to the known size of the fragments, which resulted in easier interpretation. The chosen dye label or instrument-to-instrument variations could affect the mobility of the fragments; therefore, differences in the observed length of up to ±2 bp from the expected were accepted.

2.4.3 Statistical analysis

For this section, the IBM SPSS v.22 and STATISTICA v.12 (StatSoft Inc., 2014) software were employed.

2.4.3.1 Data distribution

For each variable (e.g. age, gender or methylation ratio) the minimum, maximum, mean, median and standard deviation (SD) were calculated. If required a histogram was produced to see the shape of the distribution. Quartiles of the distribution (1st and 3rd) were also examined to assess the symmetry.

2.4.3.2 Hypothesis testing

In cases where a comparison between variables (e.g. predicted and observed age) or groups of subjects was required, an approach known as hypothesis testing was employed. Firstly, the statistical null hypothesis was set up which is the negation of the research hypothesis that generated the data. Then, the probability that the observed data were obtained if the null hypothesis were true was evaluated by calculating the p

value; the smaller the p value is the more untenable is the null hypothesis. When p values were below the cut-off of 0.05, the result was considered statistically significant (and below 0.01 it was considered highly significant), while above 0.05 it was called not significant.

2.4.3.3 Data comparison

When we had more than one groups of data, it was vital to distinguish the case where the data are paired from that where the groups are independent. With paired data the average difference between observations and the variability of these differences were assessed by paired t test. With independent data, a two sample t test was employed. In more complicated analysis, when several groups of data were obtained (e.g. several CpG sites to be analysed), an analysis of variance (ANOVA) was employed.

2.4.3.4 Linear correlation and regression analysis

The degree of linear association between two variables (e.g. methylation of a CpG site and age) was measured by calculating the correlation coefficient (r) which can take any value from -1 to +1. The correlation coefficient r measures the degree of 'straight-line' association between the values of the two variables. The linear correlation was positive if higher values of one variable were associated with higher values of the other and negative if one variable tended to be lower as the other got higher. A correlation around zero indicated that there was no linear relation between the values of the two variables (uncorrelated). Additionally, assuming that the sample was representative, a 95% confidence interval for the correlation in the population was obtained.

To evaluate if the value of one variable could be used to predict the value of the other, simple or multiple regression analysis between the outcome (dependent) variable (e.g. age) and the predictor(s) (independent) variable(s) (e.g. CpG methylation) was performed. Fitting a regression line, the sum squares of the vertical distances of the observations from the line is minimised (least squares regression); each distance is the difference between the observed value and the value given by the line, known as the fitted value or residuals. The fitted regression line explains a proportion of the variability in the dependent variable (y) and the residuals indicate the amount of unexplained variability. The proportion of the total variation explained by the model was also assessed by the goodness-of-fit of the line, which was then achieved by

considering the sum of squares explained by the regression as a percentage of the total sum of squares (R^2 value). In many cases, regression lines were linear but there were cases where the relationship between two variables was curved revealing non-linear relationships. For example, in methylation quantification by bisulphite PCR the polynomial regression was often observed either as quadratic curves ($y= ax^2 + bx + c$) or cubic curves ($y = ax^3 + bx^2 + cx + d$) as previously discussed.

2.4.3.5 Multivariate analysis

Multivariate analysis explores the association between one outcome (dependent) variable and one or more predictor (independent) variables. When studying complex concepts (for example, ageing), data can be analysed for many continuous and categorical variables. In this study, multivariate analysis used to assess if other defined factors in the datasets (such as sex) are significantly associated with the studied variable, age and create a set of these to be treated as uncorrelated variables as one approach to handling multi-collinearity as multiple regression.

2.4.3.6 Stepwise regression analysis

Stepwise regression is a semi-automated process of step-by-step iterative construction of a regression model by successively adding or removing variables based solely on the t-statistics of their estimated coefficients. It can add more power to the model, compared to the standard multiple regression option, and is especially useful for sorting through large numbers of potential independent variables (e.g. CpG sites) and fine-tuning a model by adding or removing variables. A forward stepwise regression model has been used in this study; variables are added one by one into the model, testing the addition of each variable and, adding the variable that best improves the model and repeating this process until there is no statistical improvement.

At each step, the program performs the following calculations: for each variable currently in the model, it computes the t-statistic for its estimated coefficient, squares it and reports this as its "F-to-remove" statistic, while for each variable not in the model, it computes the t-statistic that its coefficient would have if it were the next variable added, squares it and reports this as its "F-to-enter" statistic. At the next step, the program automatically enters the variable with the highest F-to-enter statistic or

removes the variable with the lowest F-to-remove statistic in accordance with certain specified control parameters.

2.4.3.7 Artificial Neural Networks (ANN) analysis

Artificial neural networks (ANNs) are a family of statistical learning algorithms inspired by biological neural networks (mainly the animal central nervous systems) and have previously been used to estimate functions that can depend on a large number of variables. Due to their adaptive nature, ANNs are generally proposed as systems of interconnected 'neurons' which can compute values from inputs and are capable of pattern recognition. They are characterised by a network topology, a connection pattern, neural activation properties and a training strategy to handle data. Statistical models are called 'neural' if they display the following properties: (a) consist of a set of adaptive weights (numerical parameters) tuned by a learning algorithm and (b) are capable of approximating non-linear functions of their inputs.

There are various types of ANNs including multi-layer perceptron (MLP), radial basis function network (RBFN), generalised regression neural network (GRNN) and probabilistic neural network (PNN). Due to notable capabilities of ANNs such as generalisation and non-linear system modelling, they have been extensively studied and applied in a range of applications (Amiri *et al.*, 2007). It is believed that ANNs are very useful in situations where the functional dependence between the inputs and outputs is not very clear. In this study, together with regression analysis, ANNs were used for building an age prediction model. It was thought that since ANN analysis allows for predicting pattern recognition, it would help in the investigation of a complex trait such as age.

In total, two different functions were applied including the MLP models examined by the STATISTICA v.12 software and the GRNN models applied through the Trajan v6 software (Trajan Software Ltd., Lincolnshire, UK). In MLPs, a neuron receives its input either from other neurons or from external inputs and has one or more hidden layers. A weighted sum of these inputs constitutes the argument of a non-linear activation function, which results in the neural output. On the other hand, GRNNs are made up of four layers of neurons including the input, pattern, summation and output layers. The main function of a GRNN is to estimate a linear or non-linear regression surface on independent variables (input vectors) given the dependent variables

(desired output vectors). Thus the network computes the most probable value of an output given only training vectors. An important benefit of GRNNs is its simplicity and fast optimisation procedure with a more efficient training process; however, it lacks a recurrent structure to filter noise. As a comparison, it has been demonstrated that the MLP models as a data modelling tool are more reliable and efficient than GRNNs when used for estimation of drug release profiles; nevertheless, when optimising drug delivery system formulation, GRNN seemed to surpass (Amiri & Derakhshandeh, 2011). The type of ANN to be used depends on the application itself.

In order to train an ANN model, the most common approach is to divide the dataset into two groups, the training and the validation (verification and blind testing) datasets. The training group is used to train the model by adjusting the weight matrices of the network model, while the validation group is used to make sure that ANN has correctly learned the relationship between inputs and outputs. The training and validation groups should contain different samples; therefore it is important that the data set is large enough; however, it is thought that ANN predicts better when using a small dataset. The selection of suitable network architecture is another significant feature as it affects the network convergence as well as the accuracy of estimations. Generally, the number of hidden neurons depends on factors like the distribution of training data and the number of samples.

It should be noted that ANN analysis was performed by Thomas Miller and Dr Leon Barron at KCL. Briefly, optimisation of each ANN type and architecture was performed in a number of stages. Firstly, all input variables (selected age-associated CpG sites) were included in the design phase followed by a second stage which enabled subset selections to assess any improved correlation or performance-limiting input variables. In these first two steps, case input variable data were randomised across training, verification and blind test experiments. When a suitable architecture was identified, a third stage of optimisation involved randomising all but the blind test cases to identify the best model. For application, all case datasets were subsequently fixed. Networks were set to balance their verification set errors against network diversity to cover as many architectures as possible across all model types. For multilayer perceptrons (MLPs) in particular, and for each set of design parameters, ANN architectures were tested over 30-minute intervals and following this, the best 50 networks were selected based on correlation and output error. Three- and four-

layer MLPs were included and within each hidden layer the number of nodes tested was ≤14 (software suggested value). For generalised regression neural networks (GRNNs), 10^8 architectures were investigated in each optimisation step. 50 of the best GRNN networks were again ranked by correlation and output error.

Part 1

3 Literature review on the identification of forensically relevant tissues

The weight of forensic scientific evidence could be enhanced in court if the cellular origin of a body fluid stain was identified. Alongside DNA typing, information regarding the cellular origin of a recovered biological stain would be very beneficial when reconstructing the events that have taken place at a crime scene. The presence of specific body fluids could be linked with particular types of crime; for example the presence of semen could indicate sexual assault. However, sometimes it is very difficult, if not impossible, to attribute a DNA profile recovered from a stain to a specific body fluid type, limiting the evidential significance of a match. Most of the current methodologies for body fluid identification use presumptive and sometimes destructive biochemical tests to identify specific elements in each fluid. Factors such as stain size or substrate material, often affect the way current methods perform limiting their capabilities. However, even in cases where a positive presumptive test is obtained, no definitive association between DNA profile and body fluid has yet been established.

The ability to accurately detect body fluids in a non-destructive manner is essential in order to protect the sample and preserve the DNA evidence. Therefore, new technologies are needed to eliminate doubts as to whether the presence of an obtained DNA profile is a truly meaningful event in the context of the case, or simply the result of innocent contact. For instance, there are a number of case examples where confirming the presence of a specific body fluid can be vital. For example, in a case where sexual abuse of a young child by a relative is suspected, the DNA profile of a person living in the same residence could be proposed to be easily deposited on the child's clothing; therefore, identifying the origin of the cells that the DNA came from (e.g. semen), would provide important probative evidence. Moreover, in a case where a sexual assault involves vaginal intercourse with a female in menses, if the victim's blood had been transferred to the suspect's clothing, the ability to identify it as menstrual in origin, and not as circulating peripheral blood, would be useful evidence that cannot be elicited using currently available presumptive tests.

3.1 Relevant background

Body fluid trace evidence found at the crime scene is one of the most significant evidence types for forensic scientists as it usually constitutes a valuable source of DNA material and can help in identifying or excluding a suspect. Detection of a body fluid usually involves a two-step procedure including a presumptive test followed by a confirmatory one. Presumptive tests are used to screen areas for suspected biological fluids, while confirmatory tests affirm the identity of a stain (Virkler & Lednev, 2009). The first step of body fluid identification is crucial as knowing the nature of each fluid is very useful per se, but the destructive nature of screening tests must be taken into account especially when only a minute amount of material is available.

3.1.1 Current presumptive testing

Generally, there are three main methods of detecting biological evidence at the crime scene: visual, microscopic, and chemical. Body fluid stains can be either visible to the eye or latent and therefore more difficult to detect. This can be due to the limited quantity of the sample, the material that the stains are deposited on (for example, dark surfaces), or because they may be masked or mixed with another body fluid. Nevertheless, some stains like semen have a characteristic colour and texture that can be easier to identify. The investigator needs to choose which test to perform first based on the physical characteristics of the stain present (Virkler & Lednev, 2009).

Various conventional immunological, serological or biochemical methods are currently used in forensic practice and a systematic biological search of the evidence is performed either at the crime scene or at the laboratory. These tests are generally quick and easy to perform, although they are generally destructive for the sample. They require large amounts of 'good quality' material present and can often lead to false positive or false negative results (An *et al.*, 2012). Several different body fluids, such as blood, semen, saliva, vaginal secretions, menstrual blood, sweat and urine, can be recovered from a crime scene, although blood, semen and saliva are the most common ones. Each body fluid has a distinctive constitution, and the presence of specific components in one fluid in comparison with another is the basis of the identification (Virkler & Lednev, 2009).

3.1.1.1 Blood

Blood is the most common body fluid recovered from crime scenes and consists of both a liquid component (plasma, 55%) and cellular fraction (erythrocytes, leucocytes and thrombocytes, 45%). Most enzymatic and chemical presumptive tests used for blood are based on the ability of the haemoglobin present in red blood cells to catalyse the oxidation of certain reagents (Matheson & Veall, 2014; Virkler & Lednev, 2009). In most cases, the oxidising agent used is a solution of hydrogen peroxide (H_2O_2). Most of these tests such as the Kastle-Mayer (KM) and the Leucomalachite Green (LGM) tests make use of reagents that change colour as a result of oxidation (Webb *et al.*, 2006). Also, luminol, known to be the most sensitive of the current presumptive tests, produces a distinct but temporary glow in a darkened environment when blood is present due to a chemiluminescent reaction (Barni *et al.*, 2007).

3.1.1.2 Semen

Semen is commonly encountered in cases of rape and sexual assault. It contains both seminal fluid and spermatozoa, even though in some cases sperm may be totally absent from the seminal fluid, a situation called azoospermia. There are a number of tests that can be used for the detection of semen stains (Virkler & Lednev, 2009). A common detection method is the observation of the suspected area under ultraviolet light in the range 250-365 nm due to the fact that semen usually fluoresces (Kobus *et al.*, 2002). However, the most popular and widely accepted presumptive test for semen is the acid phosphatase (AP) test based on this enzyme's ability to catalyse the hydrolysis of organic phosphates causing a colour change in the solution (Virkler & Lednev, 2009). Some other presumptive tests that may be applied for the detection of semen are based on the presence of semen specific proteins and other substances, such as choline (Manabe *et al.*, 1991) and seminal polyamine known as spermine (SPM) (Suzuki *et al.*, 1980).

Lastly, the most popular and definitive test for the presence of semen is the microscopical identification of sperm cells under a high-powered microscope. Spermatozoa can be easily identified by their characteristic shape after treatment with an appropriate staining reagent such as haematoxylin and eosin (Jones Jr., 2005). However, this test could give a false negative result for vasectomised men.

3.1.1.3 Saliva

Saliva is a watery (99%) fluid produced by three main pairs of salivary glands and has various functions. Saliva contains mucins, which helps with deglutition and the digestive enzyme salivary amylase (also called α amylase), which breaks down starches into maltose and dextrins. Cast-off cheek cells and bacteria cells are also present. There are only a few well-known and accepted presumptive tests for saliva, but there are no definitive tests for this body fluid. The most popular identification method is based on the activity of salivary amylase and is known as the starch-iodine test. It is based on the fact that starch appears blue when iodine is present and the salivary amylase enzyme will break down the starch to produce a colour change (Virkler & Lednev, 2009). The Phadebas® amylase test is also commonly employed; the substrate used is a water-insoluble cross-linked starch polymer carrying a blue dye, that becomes water-soluble (therefore visible) only when it is hydrolysed by α-amylase (Martin *et al.*, 2006).

3.1.1.4 Other body fluids

Since blood, semen and saliva are the most commonly recovered body fluids, efforts on the development of efficient assays have been focused on these; however, increasing research is being published on the detection of other body fluids such as vaginal fluid, menstrual blood and urine. Vaginal fluid and menstrual blood evidence can be very important in particular sexual assault cases. The constituents of these body fluids change over a female's menstrual cycle, making it very hard to develop a body fluid-specific test. As recently suggested, one solution to this problem would be to use their unique Raman spectroscopic signatures; they seem to be specific and could be used for confirmatory identification (Sikirzhytskaya *et al.*, 2012a; Sikirzhytskaya *et al.*, 2012b). Furthermore, regarding the differentiation of menstrual blood from peripheral blood, immunoassays have been proposed using monoclonal antibodies (Baker *et al.*, 2011; Gray *et al.*, 2012). Lastly, a presumptive test known as para-dimethylaminocinnamaldehyde (DMAC) has been favoured for the identification of urine as it reacts with urea; however it is not entirely specific as it results in false positives for other body fluids (Ong *et al.*, 2012).

To sum up, recovering a biological stain at the crime scene and linking it with a suspect through DNA typing is in many cases crucial for the outcome of a trial. Locating potential biological material can usually be achieved through visual examination; however, as mentioned above, presumptive testing that indicates the presence of a specific body fluid can also be employed *in situ*. Although these tests are cheap and easy to perform, it is important to understand that they are not entirely specific and require a large amount of biological material.

In more complex cases, it would be beneficial to be able to identify the cellular origin of a stain through a confirmatory test. The development of a genetic body fluid/tissue identification system that is compatible with current DNA profiling technologies would be very advantageous. Such tests have to be very sensitive, specific and maintain the integrity of DNA evidence. A panel of specific identification tests that can be performed on a single sample in a multiplex reaction is desirable, especially in cases where the sample size is minute. Protein and messenger RNA (mRNA) assays have been proven to be valuable since both are expressed in a cell-specific manner. Nevertheless, the use of proteomics - the field of multiplex analysis of complex and partially degraded protein mixtures present in stains - is relatively new and needs further improvement (Danielson, 2011; Prinz *et al.*, 2011).

3.2 mRNA profiling

Although the last decade molecular forensic science has been dominated by DNA analysis, forensic researchers have also started employing RNA technologies worldwide. The first publication dealing with post-mortem RNA analysis was in 1984 (Oehmichen & Zilles, 1984), while Phang *et al* were the first to deal with gene expression analysis (Phang *et al.*, 1994). Since then, the published work around the use of RNA in forensic science has proliferated and the forensic potential of parallel analysis is increasingly being investigated and continues to expand (Fleming & Harbison, 2010a; Hanson & Ballantyne, 2013b; Juusola & Ballantyne, 2005; Lindenbergh *et al.*, 2012). Various mRNA/DNA analysis approaches have been published, with most of them applying reverse transcription-PCR (RT-PCR) (Alvarez *et al.*, 2004; Bauer & Patzelt, 2003; Parker *et al.*, 2011). A common approach to this analysis is summarised in Figure 3-1.

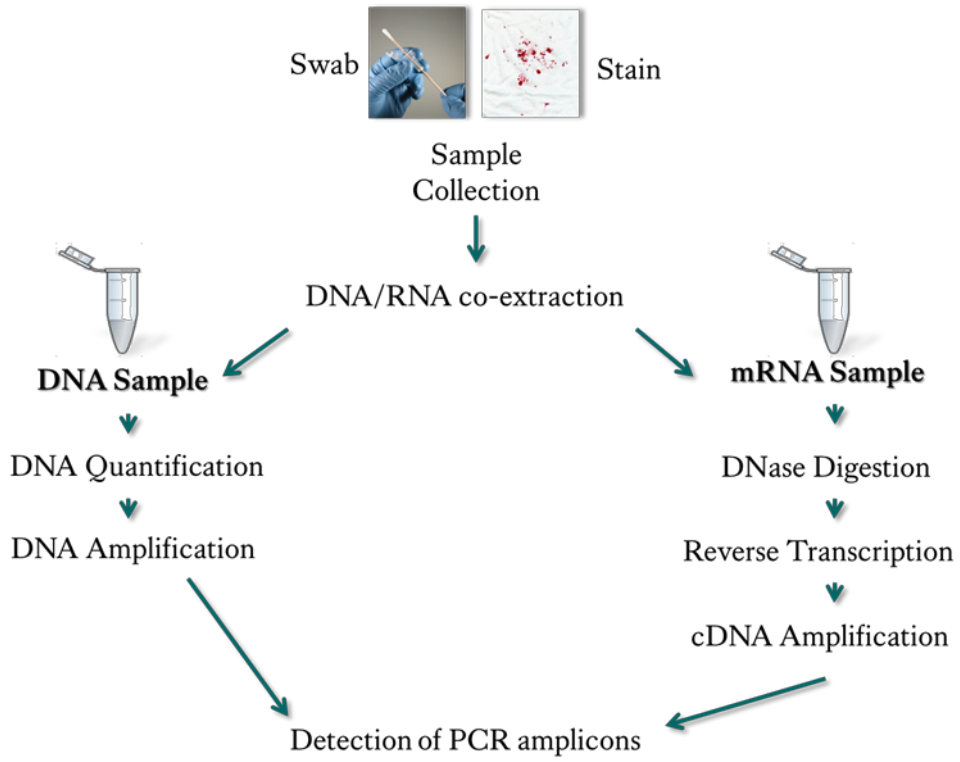


Figure 3-1. mRNA/DNA co-analysis approach
Common approach used in co-analysis of DNA and RNA in the same workflow. Recovered swabs or stains are used for simultaneous DNA/RNA co-extraction. DNA samples are analysed for standard DNA profiling, including DNA quantification, PCR and detection of STRs by capillary electrophoresis. On the other hand, the mRNA samples go through DNase treatment to eliminate potential DNA contaminants and are then used to synthesise the complementary DNA strand (cDNA) through a process called reverse transcription. Body-fluid specific mRNA markers are then amplified and detected by fragment analysis.

3.2.1 mRNA and gene expression

DNA does not direct protein synthesis itself; RNA acts as a mediator and transfers the genetic information from DNA in the nucleus to produce the protein found in the cytoplasm. The RNA molecules that are copied from genes during transcription are collectively known as messenger RNA (mRNA). It is generally accepted that RNA, compared to DNA, is highly unstable and is subject to rapid degradation by ubiquitously present ribonucleases which is essential for gene expression control (Frazao *et al.*, 2006). The final product of some genes is also RNA and these non-messenger RNAs are used in regulatory and functional processes. For example, ribosomal RNA (rRNA) is found in the centre of ribosomes and transfer RNA (tRNA) delivers amino acids during protein synthesis. There is also a class of small RNAs, microRNAs (miRNAs), whose regulatory functions in various developmental and biological procedures have been identified (Bartel, 2004).

Terminally-differentiated cells, whether they are lymphocytes (blood), spermatozoa (semen) or epithelial cells in the oral cavity (saliva), adopt a unique characteristic appearance through a regulated program, in which specific genes are turned on, while others are transcriptionally silent (Haas *et al.*, 2013b). Thus, a unique pattern of gene expression is created for each cell type, which can be demonstrated through the presence and relative abundance of specific mRNAs; a pattern known as 'transcriptome'.

3.2.2 mRNA-based body fluid identification

For the identification of forensically relevant body fluids, an mRNA-based approach would offer a range of advantages including increased sensitivity and the ability to simultaneously identify various body fluids in a semi-quantitative manner. However, to avoid false positive reactions, the selection of tissue-specific mRNA markers should be made with caution and following validation with all possible forensically relevant body fluids and tissues. Possible inter- and intra- individual variation of gene expression should also be taken into account. Lastly, it is known that single-stranded RNA is less stable than double-stranded DNA; therefore it could lead to false negative results.

Assessing the quality and quantity of RNA that can be routinely recovered from biological stains is essential in order to apply RNA-based analysis in forensic science. RNA is considered to be vulnerable to degradation by RNases and researchers initially doubted its applicability in forensic casework. However, a number of studies have showed that it is possible to detect mRNA in forensically relevant tissues (Juusola & Ballantyne, 2003) and very old/degraded stains (Kohlmeier & Schneider, 2012; Zubakov *et al.*, 2008). Zubakov *et al* managed to identify stable mRNA markers for blood and saliva identification from stains up to 16 years old. In a more comprehensive study, RNA of suitable quality and quantity could be extracted even from stains stored at room temperature for up to 547 days (Setzer *et al.*, 2008). Nevertheless, when samples were kept outdoors, this time frame was significantly reduced; up to 7 days for saliva stains; up to 30 days for blood stains and up to 180 days for vaginal secretion stains. External factors such as sunlight, humidity, high temperatures and rain have a negative influence in the stability of mRNA, which could differ among different body fluids or donors. Also, it cannot be excluded that other features such as gender, age or certain medical conditions could have an impact in mRNA patterns.

3.2.3 Tissue-specific mRNA markers

A panel of body fluid-specific mRNA markers has been identified over the past few years, which enables the identification of various forensically relevant body fluids. Although initial research had focused on developing singleplex or multiplex assays for each tissue individually, latest advances have allowed for simultaneous identification using a single method.

3.2.3.1 Blood

The mRNA-based identification of blood has been thoroughly investigated during the last decade. Various genes have been proposed to have blood-specific gene expression pattern and have been extensively validated by testing other forensically related body fluids. Initially, three mRNA candidates were broadly studied including haemoglobin beta (HBB), which is the beta-subunit of haemoglobin A (Levings & Bungert, 2002), β-spectrin (SPTB) which is an erythrocyte membrane protein (Chu *et al.*, 1994), as well as porphobilinogen deaminase (PBGD) which is an enzyme of the haeme biosynthetic pathway (Gubin & Miller, 2001). mRNA of these putative blood-specific

markers was successfully detected in peripheral blood and to a lesser degree in menstrual blood, but was undetectable in saliva, semen or vaginal secretion stains (Juusola & Ballantyne, 2005).

The results of a first collaborative European DNA Profiling group (EDNAP) exercise showed that HBB is the most abundantly-detected mRNA, SPTB showed a moderate sensitivity, while PBGD was the most difficult marker to detect (Haas *et al.*, 2011a). Some of the participating laboratories could not detect either SPTB or PBGD in any of the blood dilution series analysed; this could be due to the low level of these two mRNA species, compared to that of HBB. These findings agree with those of Patel and Peel, who reported inconsistent results for these two markers, possibly due to reduced extraction efficiency (Patel & Peel, 2008). Furthermore, researchers have tested other potential blood-specific markers such as haemoglobin alpha (HBA), which is the alpha-subunit of haemoglobin A (Nussbaumer *et al.*, 2006), erythrocyte membrane protein ankyrin (ANK1) (Fang *et al.*, 2006) as well as glycophorin A (GlycoA) (Fleming & Harbison, 2010a). Using a different approach, Zubakov and his colleagues performed whole-genome gene expression analysis on a series of time-wise degraded blood samples and reported the tissue-specific expression patterns of the most promising candidate genes by means of quantitative real-time PCR (Zubakov *et al.*, 2008). They identified a total of nine stable mRNA markers showing blood-specific expression signals in aged stains up to 180 days old. However, they reported issues with specificity with vaginal secretion due to the complex nature of this body fluid.

Haas *et al* conducted a comprehensive study analyzing 7 potential blood-specific mRNA markers including the above mentioned HBA, HBB, SPTB, PBGD and ANK1 in order to identify other mRNA markers (Haas *et al.*, 2011b). The new markers included δ-aminolevulinate synthase 2 (ALAS2) which is an erythroid-specific mitochondrial enzyme catalysing the first step of haeme biosynthesis, the CD3 gamma molecule (CD3G), a part of the T-cell receptor-CD3 complex, as well as aquaporin 9 (AQP9), a water channel protein expressed in peripheral leukocytes. While different expression levels were observed, all tested markers demonstrated great sensitivity, requiring as little as 1 ng of RNA input; however, there were some

problems with specificity (especially regarding AQP9) and tissue/animal cross-reactivity (Haas *et al.*, 2011b).

However, the use of singleplex reactions in forensic analysis is ineffective in terms of time and cost. Therefore, the above results were further validated by a second collaborative EDNAP exercise. Researchers from different laboratories tested the ability of two multiplexes to accurately identify blood: a highly-sensitive duplex consisting of both heamoglobins (HBA, HBB), as well as a moderately-sensitive pentaplex that included ALAS2, CD3G, ANK1, SPTB and PBGD (Haas *et al.*, 2012). Both multiplexes were successful enabling detection of even minute bloodstains as small as 0.01 µl, although some variability among laboratories on specificity and sensitivity were reported. Interestingly, low signals of ALAS2 have been detected in both semen and vaginal fluid stains (Richard *et al.*, 2012).

3.2.3.2 Semen

For the identification of semen, two mRNA candidates that have been widely investigated are protamine 1 (PRM1) and protamine 2 (PRM2) (Haas *et al.*, 2009b; Juusola & Ballantyne, 2005; Patel & Peel, 2008). During spermatogenesis, protamines, which are arginine-rich proteins, replace histones and become the basic nuclear proteins of mature spermatozoa (Borghol *et al.*, 2008). Both PRM1 and PRM2 have been found to be body fluid-specific, and were detected only in semen stains and absent in all other biological fluids tested (Haas *et al.*, 2009b; Juusola & Ballantyne, 2005). However, these markers were shown to also be present in some of the urine samples tested; this could presumably be due to the fact that both fluids use the same conduit (Sakurada *et al.*, 2009). The sensitivity of these markers has been tested by analysing dilution series, which could be identified from stains as little as 0.1µl (Haas *et al.*, 2009b).

Nevertheless, semen identification using protamine mRNA can only be useful when spermatozoa are present. To address the possibility of azoospermia, researchers also focused their efforts to find suitable mRNA markers within the seminal fluid. Fand and his colleagues were the first to suggest semenogelin 1 and 2 (SEMG1, SEMG2) and transglutaminase 4 (TGM4) as potential markers (Fang *et al.*, 2006). Semenogelin proteins are involved in the formation of a gel matrix that coats ejaculated spermatozoa

and are usually degraded into smaller fragments by prostate-specific antigen (Zhang *et al.*, 2008). Transglutaminases are calcium-dependent enzymes and are synthesised and secreted by prostate epithelial cells (Lu & Davies, 1997). Moreover, the mRNA of prostate-specific antigen (PSA) (or kallikrein 3, KLK3) has also been proposed as a good semen-specific candidate (Nussbaumer *et al.*, 2006). Although these mRNAs have been successfully amplified in singleplex assays (Fang *et al.*, 2006; Nussbaumer *et al.*, 2006; Richard *et al.*, 2012; Sakurada *et al.*, 2009), having one multiplex that allows for the identification of both healthy and azoospermic semen would be advantageous. Haas *et al* developed a 4-plex including PRM1, PRM2, SEMG1 and PSA mRNAs that could clearly distinguish azoospermic from normozoospermic men and obtained the expected mRNA profiles from up to 20-years-old stains (Haas *et al.*, 2009b).

3.2.3.3 Saliva

Mainly two mRNA candidates, statherin (STATH) and histatin 3 (HTN3) have been studied for saliva identification (Fleming & Harbison, 2010a; Gomes *et al.*, 2013; Juusola & Ballantyne, 2003; Patel & Peel, 2008). STATH is a stable, acidic inhibitor of the precipitation of calcium phosphate salts in the oral cavity (Sabatini *et al.*, 1990); while HTN3 is a histidine-rich protein involved in the non-immune host defence (Sabatini *et al.*, 1993). mRNAs of these saliva-specific markers were exclusively found in saliva samples and were undetectable in a variety of other body fluids, like blood and semen (Juusola & Ballantyne, 2003; Sakurada *et al.*, 2009). Both STATH and HTN3 are sufficiently sensitive and stable, as they were detected in stains as small as 0.1μl and were successfully identified in a 6-year-old stain (Sakurada *et al.*, 2009). However, some variation in gene expression for both markers has been demonstrated and also, when testing nasal secretion and vaginal fluid samples, STATH mRNA was occasionally detected, while HTN3 was undetectable (Sakurada *et al.*, 2011).

Additional potential saliva-specific genes have been suggested, such as the proline-rich proteins 1, 2, 3 (PRB1-3) (Juusola & Ballantyne, 2003) and 4 (PRB4) (Fang *et al.*, 2006). PRB proteins are known to play a role in the defence mechanisms against ingested polyhydroxylated phenols like tannins (Carlson, 1993). Additionally, as with blood, Zubakov and co-workers identified a set of five saliva-targeted mRNA markers using a microarray (Zubakov *et al.*, 2008). These include three members of the family of keratins (keratin 4 – KRT4, keratin 6A – KRT6A, keratin 13 – KRT13) known for

their role as the major structural proteins of the oral mucosa (Guo *et al.*, 2006), as well as two genes that encode for cornified envelope precursor proteins (SPRR1A, SPRR3), predominantly expressed in oral and esophageal epithelia (Gibbs *et al.*, 1993). Although all five markers were predominantly expressed in saliva, keratins were also detected to a smaller extent in semen stains (Zubakov *et al.*, 2008).

3.2.3.4 Vaginal fluid

Due to its epithelial nature, identifying vaginal fluid-specific mRNA markers is considered a difficult task, since other epithelial tissues such as the buccal epithelium and skin could have similar gene expression patterns. Nevertheless, two mRNA markers have been thoroughly investigated and showed promising results. The first is human beta-defensin 1 (HBD1) which is an antimicrobial peptide involved in the host defense of urogenital epithelium (Valore *et al.*, 1998). The second mRNA is mucin 4 (MUC4), a major membrane-spanning mucin of the endocervix that protects against pathogens and at the same time controls sperm entry into the uterus (Gipson *et al.*, 1997). Both mRNA markers are relatively specific; however, HBD1 was not always detected in vaginal fluid samples, possibly due to variations during the monthly menstrual cycle (Patel & Peel, 2008). Also, although researchers have focused on the validation of MUC4 (Juusola & Ballantyne, 2005; Nussbaumer *et al.*, 2006; Richard *et al.*, 2012), this mucin has been previously reported to be expressed also in salivary glands (Liu *et al.*, 2002).

In an effort to identify additional potential markers, estrogen receptor 1 (ESR1) was suggested for the detection of vaginal secretions as it showed >100-fold higher expression in vaginal fluid compared to other tissues (Fang *et al.*, 2006). Also, whole transcriptome profiling (RNA-seq) revealed two novel highly specific mRNA markers, namely myozenin 1 (MYOZ1) and cytochrome P450, family 2, subfamily B, polypeptide 7 pseudogene 1 (CYP2B7P1), both of which demonstrated high sensitivity (250 pg - 1 ng) and no cross-reactivity with other tissues (Hanson & Ballantyne, 2013a). Although the function of both these gene transcripts is as yet unknown, it is thought that MYOZ1 acts as an intracellular binding protein involved in linking Z-disk proteins (Takada *et al.*, 2001) and CYP2B7P1 is a non-coding RNA (pseudogene) related to the cytochrome P450 gene family involved in the oxidation of organic compounds (Laaksonen *et al.*, 1995).

Interestingly, latest research published on the identification of vaginal secretions has taken account of the body fluid's microbial signature. *Lactobacilli* species have been reported to be the main vaginal bacteria (Nam *et al.*, 2007) and closely-related species can be identified by analysing their 16S-23S rRNA intergenic spacer region (ISR) (Song *et al.*, 2000). For example, *Lactobacillus crispatus* (Lcris) and *Lactobacillus gasseri* (Lgas) have been largely investigated by the forensic community (Fleming & Harbison, 2010b; Giampaoli *et al.*, 2012; Song *et al.*, 2013). All tested vaginal swab samples showed consistent peaks for these two species; sometimes, the *Lactobacillus* mRNA peaks were higher than the ones obtained for the housekeeping genes suggesting a great number of bacteria present (Fleming & Harbison, 2010b). Other *Lactobacilli* species such as *Lactobacillus iners* (Liners) and *Lactobacillus jensenii* (Ljen) as well as other bacteria like *Gardnerella vaginalis* (Gvag) and *Atopobium vaginae* (Avag) associated with common bacterial vaginosis have also been reported to be vaginal fluid-specific (Akutsu *et al.*, 2012).

3.2.3.5 Menstrual blood

Due to its complex nature, menstrual blood is one of the most challenging body fluids. Changes throughout the menstrual cycle result in the expression of different gene transcripts; therefore, researchers have tried to identify those mRNA markers that are constantly detected regardless of the day of the cycle. The family of zinc-dependent matrix metalloproteinases (MMP) has been tested since they are specifically expressed in the endometrium during menstruation (Goffin *et al.*, 2003). The sensitivity and specificity of matrix metalloproteinase 7 (MMP7) has been supported by various research groups (Bauer & Patzelt, 2008; Patel & Peel, 2008; Richard *et al.*, 2012); MMP7 mRNA was detected in 54/60 menstrual swabs (Bauer & Patzelt, 2008).

Matrix metalloproteinases 10 and 11 (MMP10, MMP11) were also successfully detected in menstrual blood (Fleming & Harbison, 2010a; Hanson & Ballantyne, 2013b); however, they occasionally gave positive signals in vaginal fluid (Jakubowska *et al.*, 2013) or even in semen (Roeder & Haas, 2013). As expected, investigators also noticed a fluctuation of marker expression levels between women and during the menstrual cycle; MMP11 was mainly detected in the first half of the cycle (Jakubowska *et al.*, 2013). Three additional markers have also been investigated, namely mshhomeobox 1 (MSX1), secreted frizzled-related protein 4 (SERP4) and

left-right determination factor 2 (LEFTY2); however, their specificity when tested in other tissues appeared to be questionable (Roeder & Haas, 2013).

3.2.3.6 Skin

In specific forensic case scenarios, the identification of skin in touch/contact forensic samples is required. Numerous studies have emphasised that due to the minute amount of recovered biological material, it is often impossible to determine the tissue source of touch DNA evidence (Gilder *et al.*, 2009; Lowe *et al.*, 2002). Nonetheless, five novel markers were identified with the ability to detect skin from input RNA as low as 5-25 pg (Hanson *et al.*, 2011). These mRNAs belong to the late cornified envelope genes 1C, 1D and 2D (*LCE1C*, *LCE1D*, *LCE2D*), interleukin 1 family member 7 (*IL1F7*, also known as *IL37*) as well as chemokine ligand 27 (*CCL27*). The biological function of these candidates is consistent with the high specificity in skin; for example, LCE proteins are part of the epidermal differentiation complex (Jackson *et al.*, 2005), while CCL27 is responsible for recruiting cutaneous lymphocyte-associated antigen (CLA^+) memory T cells to normal or inflamed skin (Fujimoto *et al.*, 2008). Moreover, IL1F7 has a significant role in inhibiting inflammatory reactions and is mainly expressed in keratinocytes (Nold *et al.*, 2010). LCE1C and CCL27 were the most sensitive markers with a limit of detection of 5 pg RNA (Hanson *et al.*, 2011; Hanson *et al.*, 2012).

Visser and his colleagues further identified three new mRNA transcripts showing a strong over-expression in skin, which were able to successfully identify thumbprints stored for more than 6 months (Visser *et al.*, 2011). The set of markers included corneodesmosin (CDSN), loricrin (LOR) and keratin 9 (KRT9). Both CDSN and LOR take part in the assembly of the epidermal cornified cell envelope (Kalinin *et al.*, 2001), whereas KRT9 is mostly expressed in the suprabasal cells of the epidermis (Chu & Weiss, 2002). In their study, KRT9 had the lowest specificity with very low expression levels in semen and blood as well as poor sensitivity as it was not detected in all the skin samples analysed – more than 70% dropouts in quarter print samples (Visser *et al.*, 2011). On the other hand, LOR and CDSN have successfully been validated in other studies (Gomes *et al.*, 2013; Lindenbergh *et al.*, 2012).

3.2.4 Housekeeping mRNA markers

The use of housekeeping genes (HKGs) in mRNA profiling could assist in assessing the amount of mRNA recovered from a sample and potentially establishing its suitability for downstream applications. The importance of co-analysing HKGs to act as endogenous positive controls became evident when researchers were still exploring the feasibility of mRNA profiling in forensic casework. The mRNAs of ribosomal protein S15 or S18 (RPS15/RPS18), beta-actin (ACTB) and glyceraldehyde-3-phosphatedehydrogenase (GAPDH) have been used together with body fluid-specific mRNA markers (Bauer & Patzelt, 2008; Juusola & Ballantyne, 2003; Lindenbergh *et al.*, 2012). Furthermore, Fleming and Harbison (2010) suggested another three HKGs including encoding translation elongation factor-1a (TEF), glucose-6-phosphate dehydrogenase (G6PDH) and ubiquitin-conjugating enzyme E2D 2(UCE); however, TEF seemed to have an unexpected pseudogene that could interfere with interpretation (Fleming & Harbison, 2010a).

Even though many body-fluid specific markers have been identified, not enough work has been done in identifying and validating constitutively-expressed housekeeping genes that could be of forensic value and be used for data normalization. Moreno *et al* tested six HKGs: *GAPDH, ACTB* together with beta-2 microglobulin *(B2M),* phosphoglycerate kinase 1 *(PGK1),* cyclophilin A *(PPIA)* and large/acidic ribosomal protein P0 *(RPLP0)*, utilising a two-step real-time PCR assay. B2M appeared to provide the strongest overall expression pattern; however, a significant variation in expression levels for all markers was observed among different body fluids and individuals (Moreno *et al.*, 2012).

3.2.5 Multi-tissue mRNA based assays

Clearly an mRNA-based approach for body fluid/tissue identification is very promising. mRNA profiling has evolved from singleplex PCRs used to identify body fluid-specific mRNAs, to multiplex PCR assays to further validate the identified body fluid-specific transcripts. However, the development of a multiplex assay that could simultaneously detect several body fluids/tissues would be ideal and a lot of effort is currently made towards achieving this goal. The first multiplex mRNA system was proposed in 2005 and consisted of two markers per body fluid (e.g. blood, semen,

saliva and vaginal secretion) (Juusola & Ballantyne, 2005). Since then and while new research continued to reveal potential body fluid-specific mRNAs, two different combinations of markers and tissues (including also menstrual blood) were developed (Fleming & Harbison, 2010a; Patel & Peel, 2008).

Lindenbergh *et al* developed an end-point RT-PCR assay that simultaneously amplifies 19 mRNA markers, specific for blood (3), saliva (2), semen (2), menstrual secretion (2), vaginal mucosa (2) and skin (2) and allows their differentiation. The multiplex also contains three general mucosa markers as well as three housekeeping genes. Authors investigated not only the multiplex PCR assay performance and the specificity of the selected markers but also their sensitivity. Full RNA profiles were obtained with as little as 0.05 µl of starting material whereas full DNA profiles were also observed when using at least 0.1 µl. Since skin markers had not been incorporated in any multiplex system before, they were analysed in more detail. Interestingly, they resulted in successful detection of hand, feet, back and lips samples. The ability of the assay to effectively analyse old and degraded specimens up to 28 years old was remarkable (Lindenbergh *et al.*, 2012).

Furthermore, two more multiplex RT-PCR assays using different experimental approaches were published. Hanson and Ballantyne (2013) proposed a hexaplex high-resolution melt (HRM) protocol in order to simplify the mRNA profiling process and also reduce the time and cost of the analysis. In their assay, they included one marker per body fluid; ALAS2 for blood, HTN3 for saliva, TGM4 for semen, IL19 for vaginal secretion, MMP10 for menstrual blood and IL1F7 for skin identification. The assay showed high sensitivity (pg of DNA) and no significant cross-reactivity (Hanson & Ballantyne, 2013b). Additionally, a multiplex quantitative RT-PCR assay including four novel markers – PPBP for blood, FDCSP for saliva, MSMB for semen and MSLN for vaginal secretion – was studied (Park *et al.*, 2013). The selected markers showed good specificity and sensitivity, though the MSNL was less optimal in sensitivity as previously expected.

Table 3-1 summarises the main body fluid-specific markers reported in literature for blood, saliva, semen, vaginal secretion and menstrual blood.

Table 3-1. mRNAs found in the literature showing body fluid-specific expression

Body fluids	Body-fluid specific mRNAs					
Blood	HBA	Haemoglobin α	MNDA	Myeloid nuclear differentiation antigen	AQP9	Aquaporin 1
	HBB	Haemoglobin ß	AMICA1	Adhesion molecule	ANK1	Ankyrin 1
	SPTB	ß-spectrin	CD88	Cluster of differentiation 88	GlycoA	Glycophorin A
	PBGD	Porphobilinogen deaminase	CD93	Cluster of differentiation 93	CD3G	CD3 γmolecule
	ALAS2	δ-aminolevulinate synthase 2	PPBP	Pro-platelet basic protein	CASP1	Caspase 1
Saliva	STATH	Statherin	PRB3	Proline-rich protein 3	KRT13	Keratin 13
	HTN3	Histatin 3	PRB4	Proline-rich protein 4	KRT6A	Keratin 6A
	MUC7	Mucin 7	SPRR3	Small proline-rich protein 3	KRT4	Keratin 4
	PRB1	Proline-rich protein 1	SPRR1A	Small proline-rich protein 1A		
	PRB2	Proline-rich protein 2	FDCSP	Follicular dendritic cell secreted protein		
Semen	PRM1	Protamine 1	PSA	Prostate-specific antigen		
	PRM2	Protamine 2	MCSP	Mitochondria cysteine-rich protein		
	SEMG1	Semenogelin 1	MSMB	Beta-microseminoprotein		
	SEMG2	Semenogelin 2				
	TGM4	Transglutaminase 4				
Vaginal fluid	HBD1	Human beta-defensin 1	IL19	Interleukin 19	Ljen	16S rRNA
	MUC4	Mucin 4	FUT6	Fucosyltransferase 6	Lcris	16S-23S rRNA
	ESR1	Estrogen receptor 1	SFTA2	Surfactant associated 2	Lgas	16S-23S rRNA
	MYOZ1	Myozenin 1	DKK4	Dickkopf homolog 4	Liners	16S-23S rRNA
	CYP2B7P1	CytP450, fam2, subB, pol7, pseud1	MSLN	Mesothelin	Avag	16S rRNA
Menstrual blood	MMP7	Matrix metalloproteinase 7	SFRP4	Secreted frizzled-related protein 4		
	MMP10	Matrix metalloproteinase 10				
	MMP11	Matrix metalloproteinase 11				
	MSX1	Mshhomeobox 1				
	LEFTY2	Left-right determination factor 2				

3.2.6 Interpretation of mRNA profiles

As shown above, there is substantial evidence to support that mRNA profiling could serve as confirmatory testing to identify the cellular tissue of origin. However, significant variation in terms of specificity and sensitivity were observed across research groups. Also, the majority of published PCR multiplex assays use one or two markers for the identification of each body fluid which often leads to false positive results. It has been proposed that a minimum of five markers per body fluid should be included (Roeder & Haas, 2013).

Also, given the 'occasional' expression of some markers in non-target tissues and the observed intra- and inter- individual variation (especially for vaginal fluid and menstrual blood markers), it is essential that a scoring system is developed to set the standards for the interpretation of mRNA profiles. Recently, the Netherland Forensic Institute published their guidelines on how mRNA profiling was implemented in forensic casework (Lindenbergh *et al.*, 2013). They developed a stepwise procedure that could prevent cognitive (confirmation, expectation or motivational) bias:

- The researcher who performs the mRNA profiling should remain uninformed about the context of the case.
- mRNA results are generated according to specific guidelines and are presented in the form of a table with six different scoring categories.
- DNA results are generated separately following standard routine analysis.
- DNA/mRNA profiles are interpreted and a relationship between individuals and body fluids detected is established.
- Results are collated and conclusions are formulated in a final report.

After analysing specially-designed mock cases, the authors were able to show that potential bias was eliminated and that results generated from presumptive testing, mRNA profiling and DNA analysis were concordant (Lindenbergh *et al.*, 2013). However, one has to be careful when associating cell types and donors as variations in both DNA and RNA stability have been reported (Harteveld *et al.*, 2013).

As shown above, mRNA profiling using tissue-specific markers can be considered as a good confirmatory test as most of the proposed mRNA markers are very sensitive giving a positive reaction with as low as 0.05 µl stain (Lindenbergh *et al.*, 2012). Selected markers can be multiplexed allowing for simultaneous identification of various forensically relevant tissues. However, there are issues regarding tissue-to-tissue specificity and also when applying these tests in extensively-degraded samples where RNA is more difficult to obtain. Such constraints are particularly important when re-examining 'cold cases', where only DNA has been retained and current methods cannot be used. A method that would not consume additional sample during nucleic acids co-extraction and at the same time is able to exploit the stability of DNA molecule would be preferred. DNA-based tests could potentially overcome the limitations of existing methods and provide a direct link between the recovered DNA and its source.

It is known that DNA methylation is one of the main mechanisms responsible for cell differentiation and differential gene expression (Plachot & Lelievre, 2004; Song *et al.*, 2009). As previously described in section 1.1.2, DNA methylation regulates gene expression by mostly silencing (or in some cases activating) gene transcription. Therefore, differential DNA methylation patterns could offer a great alternative to mRNA profiling when differentiating forensically relevant tissues.

3.3 Tissue-specific DNA methylation

Over the last decade, tissue-specific methylation patterns have been investigated either by analysing specific gene loci (Illingworth *et al.*, 2008) or entire chromosomes (De Bustos *et al.*, 2009). There are numerous studies that have looked at differential methylation patterns across various tissues such as between blood and muscle (Liang *et al.*, 2008) or between blood and saliva (Thompson *et al.*, 2013). Moreover, researchers have investigated how these tissue-specific patterns change over the individual's lifetime (Christensen *et al.*, 2009). Rakyan and co-workers reported a novel resource for human genome-wide tissue-specific DNA methylation profiles across 13 normal somatic tissues, placenta, sperm and an immortalized cell line (Rakyan *et al.*, 2008). Authors used this resource to identify the first comprehensive genome-wide set of tissue-specific differentially methylated regions (tDMRs) that could play a role in cellular identity as well as the regulation of tissue-specific genome function. DNA methylation analysis has already been proposed as a tool for cell typing in cell therapeutic approaches. As an example, Baron *et al* identified panels of cell type-specific differentially methylated gene regions (CDMs) using a methylation-sensitive single-nucleotide primer extension (MS-SNuPE) protocol, which allowed for accurate identification and quantification of subpopulations in cell cultures (Baron *et al.*, 2006).

3.3.1 DNA methylation-based tissue identification

From a forensic point of view, scientists thought that DNA methylation profiling could potentially be very helpful in confirming the presence of a specific tissue. Differentiation could be achieved via the identification of suitable DNA methylation markers that are differentially methylated amongst forensically relevant tissues. A DNA methylation-based approach would potentially overcome limitations of existing methods including the additional consumption of evidence or qualitative rather than quantitative findings, but also offer further advantages. Such tests are more likely to demonstrate great sensitivity and robustness since DNA is very stable.

Frumkin *et al* were the first to explore the possibility of DNA methylation-based forensic tissue identification (Frumkin *et al.*, 2011). 205 individual CpG islands containing a recognition sequence for the methylation-sensitive restriction enzyme Hha I were selected for initial screening for differential methylation by comparing

signals obtained from pooled DNA samples (10 individuals) for each tissue. As a result, a total of 38 genomic loci demonstrated differential amplification patterns, 15 of which were used in their tissue identification assay [Table 3-2]. To develop and validate the assay, 1 ng of 50 DNA samples including blood, saliva, semen, skin, urine, menstrual blood and vaginal secretion were analysed; Figure 3-2 presents an overview of their approach. Since highly methylated loci were protected from enzymatic digest and amplified with much higher efficiency than loci with lower methylation levels, the authors considered only ratios of methylation levels in their analysis. Even though variability in methylation ratios due to inter-individual variation or stochastic PCR effects were observed, authors suggested that each tissue type had a distinct methylation profile. The proposed assay was accompanied by an in-house developed algorithm that could correctly report the 'true' tissue type for all 50 samples. Further analysis also revealed that 100% identification could be achieved by using only seven out of the 15 loci. The suggested method could be very promising as a forensic application since it utilises the same platform used in standard STR profiling; however, incomplete digestion by the methylation-sensitive restriction enzyme could lead to erroneous results. Furthermore, methylated loci not amplified because of degradation in aged samples could be mistaken to be unmethylated. Interestingly, Gomes *et al* failed to reproduce their results on skin identification since they observed the same methylation profile for skin and saliva (Gomes *et al.*, 2011). Consequently, the suggested loci require further investigation for possible inter-individual variation.

Table 3-2. Proposed tissue-specific CpG sites (Frumkin *et al.*, 2011)

CpG sites	Locus	Chromosomal location
1	L91762	12: 73,985,697
2	L68346	3: 12,175,317
3	L50468	3: 55,492,965
4	L14432	22: 29,866,149
5	L4648	1: 35,815,331
6	L39664	17: 53,710,330
7	L30139	17: 36,996,489
8	L55429	5: 1,548,043
9	L62086	19: 40,478,538
10	L76138	19: 3,130,217
11	L15952	7: 2,741,305
12	L36599	19: 4,867,642
13	L26688	17: 77,844,826
14	L81528	19: 47,395,603
15	L36556	19: 50,962,118

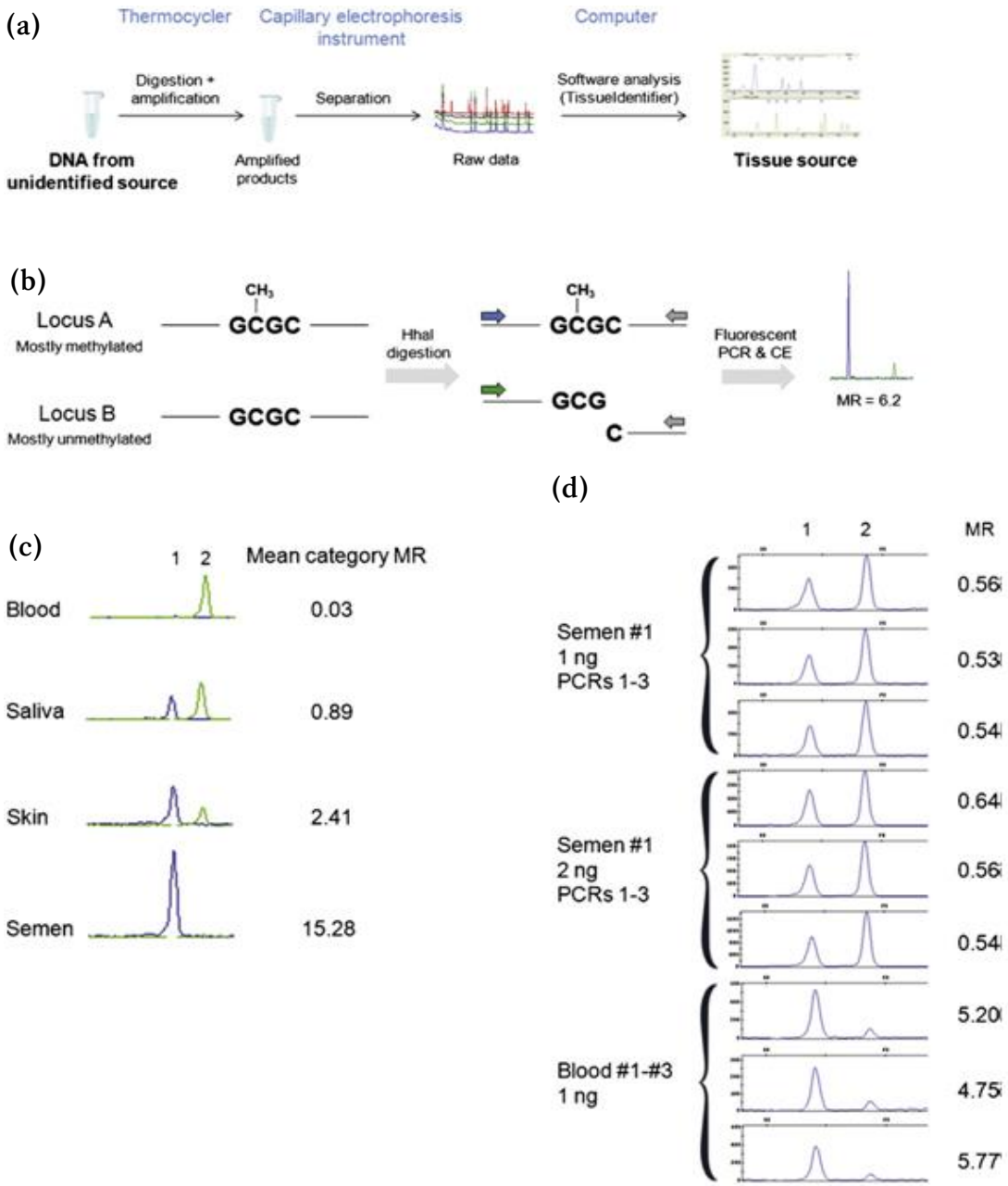


Figure 3-2. Overview of the tissue identification assay (Frumkin *et al.*, 2011)

(a) Description of methodology used, (b) During digestion with methylation-sensitive restriction enzyme highly methylated loci remain intact and subsequently are amplified during PCR (locus A), whereas low-methylated loci are mostly digested resulting in a weak signal due to inefficient PCR amplification (locus B); the methylation ratio (MR) corresponds to the differential methylation level between these two loci, (c) Obtained differential methylation ratios in blood, saliva, skin and semen, (d) Observed methylation differences between blood and semen samples are greater than differences observed due to different PCRs or starting DNA amount.

Additionally, Lee *et al* tested previously reported, tissue-specific differentially methylated regions (tDMRs) in the literature and proposed a different methodological approach that included bisulphite sequencing (Lee *et al.*, 2012a). After testing pooled DNA from blood, semen, saliva, menstrual blood and vaginal fluid, the authors confirmed two previously reported testis-specific DMRs, namely dapper 1 isoform 2

(*DACT1)* and ubiquitin carboxyl-terminal hydrolase 49 *(USP49*) that could potentially be applied for semen identification. They also tested another three tDMRs belonging to the homeobox protein Hox-A4 (*HOXA4),* profiling-3 *(PFN3)* and protein arginine N-methyltransferase *2 (PRMT2*) genes which had previously displayed varying degrees of methylation among blood, brain, muscle and spleen. As the authors suggested, the presence of an unmethylated *HOXA4* tDMR could possibly be used to exclude the presence of blood and the *PFN3* tDMR could potentially be used for the identification of vaginal secretions. Although the results were promising, sex differences and inter-individual variations were once again observed.

To further validate these tDMRs, the authors employed two more sensitive multiplex assay systems using four out of the five proposed markers (*DACT1, USP49, PFN3* and *PRMT2*): a methylation-specific restriction enzyme PCR (MSRE-PCR) technique and a methylation SNaPshot protocol (An *et al.*, 2013). They investigated age-associated methylation changes by analysing body fluids from both young and elderly men and reported no significant differences. Both systems were able to successfully identify semen and differentiate menstrual blood and vaginal fluid from blood and saliva samples. However, the discrimination power regarding the non-semen body fluids was increased when they integrated selected body fluid-specific microbial DNA markers and a previously reported semen-specific marker by Frumkin and co-workers (*L81528,* Table 3-2). As a result, the identification of saliva was achieved by detecting saliva-specific bacteria whereas vaginal fluid was differentiated from other body fluids via *Lactobacillus* species (Choi *et al.*, 2013).

Employing a similar approach, Wasserstrom *et al* developed a DNA methylation-based semen test (Nucleix DSI-Semen kit), which could successfully distinguish between semen and non-semen samples and could potentially replace the time-consuming microscopic examination of casework samples (Wasserstrom *et al.*, 2013). A panel of five genomic loci that were believed to demonstrate substantial DNA methylation differences between semen and all other body fluids tested were selected and tested on 135 DNA samples. The accuracy of the kit was high and the accompanying software also developed by the authors made the subsequent data analysis more reliable. LaRue *et al* performed a more comprehensive validation study on the kit's performance and illustrated that the required starting DNA material can be as low as 62 pg (LaRue *et*

al., 2013). Nevertheless, a test that could simultaneously identify all body fluids would be preferable.

Most methods mentioned above use a restriction enzyme-based qualitative approach rather than provide accurate quantitative methylation results. Following a similar protocol, developed by Paliwal and co-workers which allowed for quantitative detection of DNA methylation states in minute amounts of DNA from body fluids (Paliwal *et al.*, 2010), Madi *et al* proposed a sensitive method that could be applied in trace levels of forensic samples (Madi *et al.*, 2012). Utilising bisulphite Pyrosequencing®, they could accurately evaluate the relative quantity of methylated cytosines of various adjacent CpG sites at four different genomic loci. The encoding chromosome 20 open reading frame 117 (*C20orf117)* region was found to be blood-specific, while zinc finger CCCH-Type Containing 12D (*ZC3H12D)* and fibroblast growth factor 7 (*FGF7)* showed sperm-specific methylation. Lastly, the region of the breast carcinoma amplified sequence 4 (*BCAS4)* seemed to be saliva-specific although it had initially been proposed as a semen-specific marker (Eckhardt *et al.*, 2006). It is, however, notable that methylation differences between tissues were not always high which could also complicate the analysis of mixed body fluid stains [Figure 3-3]. Other forensically relevant body fluid types should also be investigated to establish the marker specificity.

Moreover, Ma *et al* screened six potentially blood-specific tDMRs using methylation-sensitive difference analysis and Sequenom Massarray utilising methylation-sensitive representational difference analysis (MS-RDA) technology (Ma *et al.*, 2013). The authors identified two fragments showing blood-specific hypomethylation and four fragments showing blood-specific hypermethylation; however no menstrual blood samples were tested so it is unclear if these tDMRs are only peripheral blood-specific.

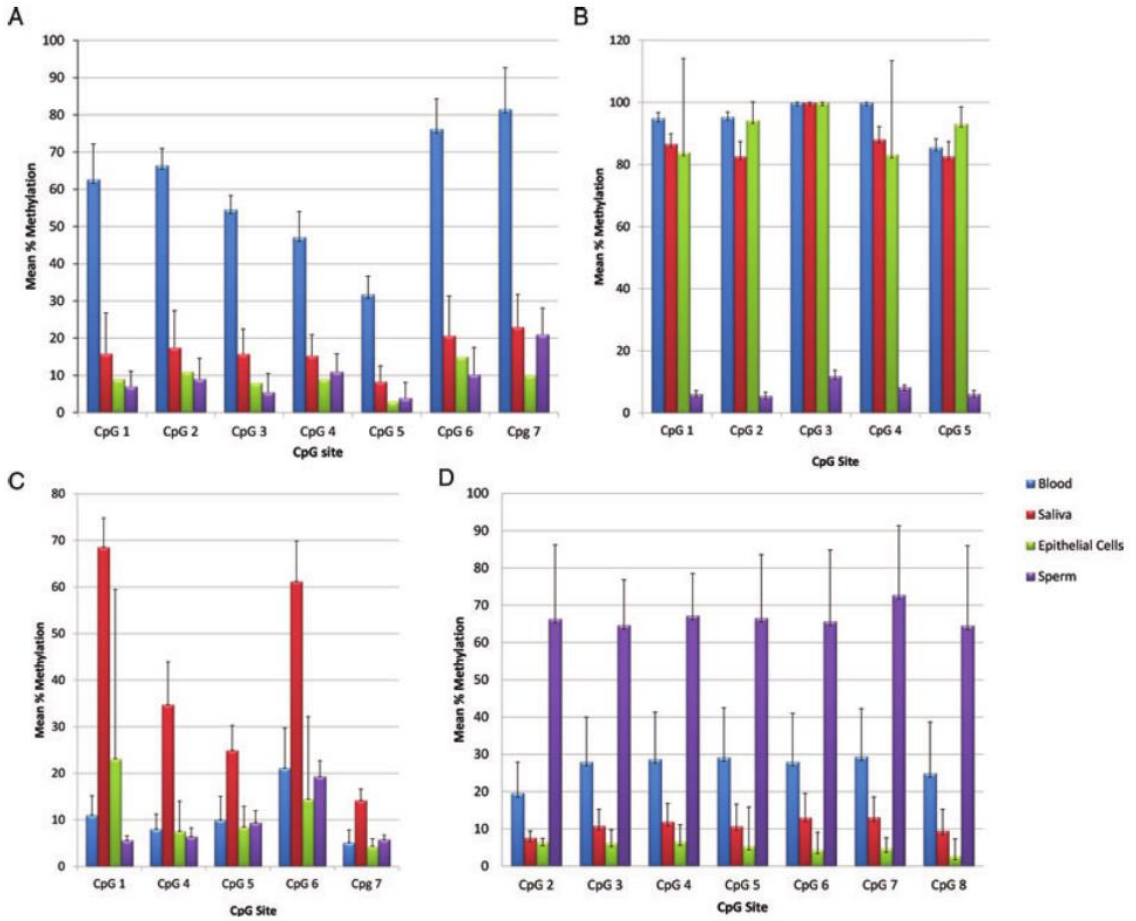


Figure 3-3. Mean methylation values obtained for all analysed genomic loci following analysis of four different tissue types (Madi *et al.*, 2012)

(A) *C20orf117* (7 CpGs), (B) *ZC3H12D* (5 CpGs), (C) *BCAS4* (5 CpGs) and (D) *FGF7* (7 CpGs)

3.4 Conclusion

As shown in this review, the identification of forensically relevant body fluids in a confirmatory way can be quite challenging. Although efforts towards determining tissue-specific mRNA and DNA methylation markers have been made, factors such as natural inter-individual variation in gene expression or DNA methylation affect marker specificity. As observed in most studies, using either mRNA or epigenetic markers the detection of biologically complex body fluids containing more than one cell type (e.g. vaginal fluid) has been proved to be particularly difficult. Regarding mRNA profiling, multiplex PCR systems that allow for simultaneous identification of various tissues has already been developed; however, further validation via developing an appropriate analysis strategy and establishing interpretation guidelines is required. On the other hand, with the exception of semen, determining highly tissue-specific differentially methylated CpG sites for the rest of the analysed tissues still remains a challenge. Studies so far show promising results, with several working at picogram levels of DNA which is necessary for forensic applications; however, it is important to either discover more tissue-specific CpG sites or further validate previously reported ones in the literature.

4 Evaluation of multiplex tissue-specific mRNA-based systems

4.1 Introduction

As mentioned earlier in section 3.2.3.1, two collaborative exercises on mRNA profiling for the identification of blood were organised by the European DNA Profiling (EDNAP) group in order to evaluate the robustness and reproducibility of the proposed mRNA markers (Haas *et al.*, 2012; Haas *et al.*, 2011a). Various forensic genetic laboratories across Europe analysed a set of samples using kits and chemistries of their own choice as well as their own instrumentation. Since most of the participating laboratories had no experience with RNA prior to the first collaborative exercise, the main goal was to implement the method and compare the sensitivities of the markers.

On the same principle, the Institute of Legal Medicine in the University of Zürich, Switzerland organised another four collaborative exercises on behalf of the EDNAP Group regarding the identification of semen, saliva, vaginal secretion, menstrual blood and skin:

- The third collaborative EDNAP exercise on RNA/DNA co-analysis was related to the identification of semen and saliva. It involved testing a saliva triplex including HTN3, STATH and MUC7 mRNA markers and a novel semen pentaplex allowing for the detection and differentiation of both sperm (PRM1, PRM2) and seminal plasma (PSA, SEMG1, TGM4). Therefore, semen stains obtained from both healthy and azoospermic men could be identified. In addition to the multiplexes, four singleplex PCRs (HTN3, MUC7, PRM2 and SEMG1) were analysed permitting a comparison of the results obtained using singleplex and multiplex analysis.
- The fourth EDNAP exercise was related to the identification of menstrual blood. It involved testing two menstrual blood multiplexes; the MMP triplex including MMP7, MMP10, MMP11 mRNA markers and the MB triplex including the markers MSX1, LEFTY2 and SFRP4.
- The fifth EDNAP exercise was relevant to the identification of vaginal secretion. It involved testing two vaginal fluid multiplexes; the human Vag triplex consisted of MYOZ1, CYP2B7P1 and MUC4 mRNAs and the bacterial Lacto triplex containing the markers Ljen, Lcris and Lgas. Together with the multiplexes, a HBD1 singleplex was analysed.

- Lastly, the sixth EDNAP exercise was organised to test suitable markers for the identification of skin cell deposits and included two multiplexes as well; Skin1 co-amplifying five mRNA markers (LCE1C, LCE1D, LCE2D, IL1F7 and CCL27) and Skin2 amplifying LOR, KRT9 and CDSN mRNA transcripts. Additionally, a housekeeping (HKG) triplex was proposed that included the B2M, UBC and UCE markers.

Furthermore, as part of the European Forensic Genetics Network of Excellence (EuroForGen), another collaborative mRNA exercise was organised which involved testing a 'corrected' version of the multi-tissue mRNA system proposed by Lindenbergh *et al.* (2012). This adjusted protocol amplifies 20 mRNA markers in total and is able to differentiate between blood, saliva, semen, vaginal secretion, menstrual blood and skin. It should be noted that it is the only multiplex mRNA system available so far that allows for simultaneous identification of all body fluids and is currently being implemented in forensic casework in the Netherlands.

4.1.1 Aim and Objectives

The aim of this study was to evaluate and assess the performance of several multiplex mRNA systems as part of collaborative mRNA profiling projects.

For each validation exercise, a set of blind samples was analysed. The sensitivity, specificity and applicability of each multiplex RT-PCR method were assessed. Lastly, the proposed interpretation guidelines were also implemented using mock casework samples.

4.2 Experimental

Each EDNAP/EuroForGen exercise was accompanied with general remarks and recommendations on methods and instrumentation settings. The methodology employed in this study regarding DNA/RNA extraction, DNase treatment, cDNA synthesis, PCR, PCR product purification and fragment analysis via capillary electrophoresis is extensively described in the '2.3 RNA analysis' section of Chapter 2.

4.2.1 Samples

All body fluid samples together with the primers were sent on dry ice via post and stored at -20 °C prior to analysis. No information regarding the samples was revealed until the end of analysis. The whole swab or stain was used for DNA/RNA extraction, if not stated otherwise.

Regarding the EDNAP exercises:

4.2.1.1 Dilution series

Dilution series for saliva (10, 5, 1, 0.5, 0.1, 0.05 µl), semen (5, 1, 0.5, 0.1, 0.05, 0.01 µl), menstrual blood (1/4, 1/8, 1/16, 1/32, 1/64 of a menstrual blood swab), vaginal secretion (1/4, 1/8, 1/16, 1/32, 1/64 of a vaginal secretion swab) and skin RNA (200, 50, 12, 3, 0.8 ng) were analysed with the respective single- or multiplexes in order to test the sensitivity of the markers. Specifically for the housekeeping triplex, three dilution series of blood (1, 0.1, 0.01, 0.001 µl), semen (5, 1, 0.2, 0.04 µl) and saliva (25, 5, 1, 0.2 µl) were tested. For the dilution series, liquid samples were diluted in 0.9% NaCl to a final volume of 5 µl per sample and were then deposited on cotton swabs. Menstrual blood and vaginal secretion swabs were collected from healthy females while commercial skin RNA was acquired (Agilent Technologies). All swabs were left to dry at room temperature overnight.

4.2.1.2 Body fluid stains

Moreover, each exercise included a blind set of stains consisting of either the tissue of interest, non-target tissues, age/degraded stains or non-human samples in order to test the specificity of the markers as well as the applicability in forensic casework. The analysed stains (34 in total) are presented in Table 4-1.

Table 4-1. EDNAP mock casework samples consisted of different body fluids

Stains		Description
Set 1	1	5 µl of saliva on paper
	2	3 µl of azoospermic semen on pad
	3	¼ of a vaginal swab
	4	5 µl of saliva on a glass slide
	5	2 µl of semen on toilet paper
	6	1 µl of semen/5 µl of saliva on a swab
	7	¼ of a buccal swab (cat)
	8	1 µl of blood/2 µl of saliva on a swab
	9	Plastic spoon
	10	3 µl of semen inside latex glove
Set 2	11	Menstrual blood on sanitary towel (fresh)
	12	¼ of a cotton swab with EDTA-blood (fresh)
	13	¼ of a menstrual blood swab (fresh)
	14	Menstrual blood on sanitary towel (5 years old)
	15	¼ of a vaginal swab (fresh)
	16	¼ of a menstrual blood swab (5 years old, day 1-4)
	17	¼ of a menstrual blood swab (5 years old, day 1/day 4)
	18	¼ of a menstrual blood swab (fresh)
Set 3	19	¼ of a vaginal swab (2 years old)
	20	5x5 mm from white worn underpants (fresh)
	21	½ of a swab with urine (fresh)
	22	½ of a vaginal swab, pregnant (fresh)
	23	¼ of a vaginal swab (5 years old)
	24	½ of a vaginal swab (fresh)
	25	½ of a buccal swab (fresh)
	26	5x5 mm from sanitary towel (fresh)
Set 4	27	Small swab from palm
	28	Hand print on paper, glossy side
	29	Key from computer keyboard
	30	Fingerprint on a glass slide
	31	Small swab with urine
	32	Swab from palm/1 µl of blood
	33	Small swab with saliva
	34	Scraped skin from the back of the hand

- For set 1 (semen/saliva identification), freshly collected saliva from ten individuals was used to make up the stains by depositing it onto different surfaces like paper and glass slides. Licked plastic spoons were also provided. Previously frozen (for up to 25 years) semen from nine individuals was used to make up the semen stains onto various materials such as pad, toilet paper and latex glove. Also, blood from two donors, vaginal swab from a healthy female and buccal swab from a cat were used as the non-saliva/semen and non-human samples. All stains were prepared and left to dry at room temperature overnight.
- For set 2 (menstrual blood identification), fresh or stored (for up to 5 years at room temperature) menstrual blood from six volunteers was collected on swabs or on sanitary towels to make up the different stains. Non-menstrual blood samples were collected from three healthy females, while semen from one donor (previously frozen for up to 5 years) was used. Stain 12 comprised of half of a swab with EDTA-blood, which corresponds to around 10 µl in volume.
- For set 3 (vaginal secretion identification), fresh or stored (for up to 5 years) vaginal secretion was collected from six females either on swabs, sanitary towels or underwear. Fresh non-vaginal fluid samples were collected from two volunteers in the form of a half of a urine swab (stain 21) and half of a buccal swab (stain 25).
- For set 4 (skin identification), fresh skin samples from three different individuals were used to prepare stains 27, 28, 30, 32 and 34, while stain 29 was a key from a used computer keyboard. Non-skin stains (urine, saliva and blood) were provided by healthy donors.

4.2.1.3 In-house samples

To further validate the proposed RT-PCR assays, each laboratory tested their own target/non-target body fluid stains; therefore, additional casework/mock casework samples were prepared and tested together with the blind sets. Together with set 2, two additional stains were analysed including a menstrual blood swab (day 3) and a suspected saliva stain (unknown volume, criminal sample). Together with set 3, four additional stains were included in the analysis including a vaginal swab (day 16), half of a menstrual blood swab (day 4), a skin swab from palm as well as a 5µl saliva stain. Together with set 4, four extra samples were analysed including a skin swab from palm, a 5 µl saliva stain, a vaginal swab (day 13) and a hair (including the root).

Regarding the EuroForGen exercise:

4.2.1.4 cDNAs and purified PCR products

Purified PCR products produced by amplifying six different cell types, namely blood, semen, saliva, vaginal fluid, menstrual blood and skin were firstly analysed by all laboratories in order to adjust provided bin sets and assess any differences in CE sensitivity. Then, triplicates of single-source synthesised cDNA samples derived from different RNA inputs were analysed in order to evaluate the overall marker sensitivity. These cDNA samples had been prepared by initially extracting 10 µl of blood or semen, 20 µl of saliva, a stub of skin or a swab containing menstrual or vaginal material. Skin samples were collected using a tape lift stub (1*1 cm), until the stub was saturated. The entire stub or swab was processed for extraction; however, the exact amount of material on them was not measured by the organising laboratory. All RNA extracts were then eluted in 60 µl of RNase-free water and were not quantified. Especially for menstrual and vaginal samples, RNA inputs in cDNA synthesis were also adjusted to prevent profile overload.

Together with the above cell type-indicated cDNAs, a set of ten unknown cDNAs (including one mixture) were included in the study in order to achieve familiarisation with the serial cDNA input approach and assess overall performance of the 20-plex (marker drop-in and drop-out). These cDNAs were derived from the following samples following RNA elution in 60 µl (the utilised RNA amount into cDNA synthesis is indicated in brackets):

- **Stain 1** – a stub of skin (1 µl RNA)
- **Stain 2** – 10 µl of semen (fertile) (0.5 µl RNA)
- **Stain 3** – a menstrual blood swab (1 µl of 1:10 diluted RNA)
- **Stain 4** – 20 µl of saliva (1 µl RNA)
- **Stain 5** – No template/blank
- **Stain 6** – a vaginal swab (1 µl of 1:150 diluted RNA)
- **Stain 7** – 10 µl of semen (sterile) (1 µl RNA)
- **Stain 8** – 10 µl of blood (0.75 µl RNA)/a vaginal swab (1 µl of 1:75 diluted RNA)
- **Stain 9** – 10 µl of blood (1 µl RNA)
- **Stain 10** – a stub of skin (1 µl RNA)

These results were used to determine the optimal cDNA input ranging from 0.5 to 2 µl. Lastly, to test the usefulness of the scoring system, 4 mixed cDNAs were tested [cDNA 1 – blood : saliva (1:1), cDNA 2 – blood : saliva : skin (1:1:1), cDNA 3 – vaginal : semen (fertile) (10:1), cDNA 4 – menstrual blood : semen (sterile) : skin (10:5:1)].

4.2.1.5 Complex mock casework samples

In order to further test the utility of the proposed system for scoring mRNA profiles and its applicability in complex forensic casework, four challenging mock casework samples were provided:

- **Stain 1** – swab containing equal amounts of saliva from two donors
- **Stain 2** – piece of fleece containing a menstrual blood/whole blood mixed stain
- **Stain 3** – patch of linen with skin from two donors plus whole blood from one
- **Stain 4** – nail clipping containing skin, vaginal mucosa and azoospermic semen

To make up these stains, saliva and semen were collected in tubes, whole blood using the finger prick method, vaginal fluid on cotton swabs, menstrual secretion on Viba brushes (Rovers) and skin by rubbing textiles over the face.

4.2.2 Multiplex and singleplex RT-PCR assays

For this study, the amplification set up and conditions were followed as described below. Primer sequences and amplicon sizes are presented in Tables 4-2 and 4-3. Primers are labelled with commonly used dyes to facilitate correct spectral calibration.

Table 4-2. Primer sequences and amplicon sizes of the mRNA markers used for the identification of blood, saliva, semen, vaginal secretion, menstrual blood and skin as well as the HKG triplex (EDNAP mRNA profiling exercises)

Body fluid		Gene	Accession No.		Primer sequence (5'→3')	Dye	Size (bp)
Saliva	HTN3	Histatin 3	NM_000200	F	GCAAAGAGACATCATGGGTA	NED	134
				R	GCCAGTCAAACCTCCATAATC		
	STATH	Statherin	NM_003154	F	TTTGCCTTCATCTTGGCTCT		93
				R	CCCATAACCGAATCTTCCAA		
	MUC7	Mucin 7	NM_001145006.1	F	CTAAAAGCAAGCAACTGGAT		197
				R	AAGTGAGATTTGGGTGATTG		
Semen	PRM1	Protamine 1	Z46940	F	GCCAGGTACAGATGCTGTCGCAG	VIC	153
				R	TTAGTGTCTTCTACATCTCGGTCT		
	PRM2	Protamine 2	NM_002762.2	F	GGCGCAAAAGACGCTCC		91
				R	GCCCAGGAAGCTTAGTGCC		
	PSA	Prostate-specific antigen	NM_001648	F	TGTCCGTGACGTGGATTG		82
				R	GGTTGGGAATGCTTCTCG		
	SEMG1	Semenogelin 1	NM_198139	F	TCGGTAACCATGTGAAAGGA		120
				R	GTGTCATCCATGGACCAAGA		
	TGM4	Transglutaminase 4	NM_003241.3	F	TGAGAAAGGCCAGGGCG		215
				R	AATCGAAGCCTGTCACACTGC		
Skin	LCE1C	Late cornified envelope IC	NM_178351	F	GCTGAAGGACCCTGTGCT	6-FAM	56/58
				R	CAGGACATCTTGGTGGCG		
	IL1F7	Interleukin 1 family member 7	NM_173203	F	CCAGTGCTGCTTAGAAGACC		92
				R	TCACCTTTGGACTTGTGTGAA		
	LCE1D	Late cornified envelope ID	NM_178352	F	CCTGTGCTGCCTGTGACT		142
				R	GGCACTTAGGGGGACATTTA		
	LCE2D	Late cornified envelope 2D	NM_178430	F	TCTGTGCTTTTGCATGTGAC		193
				R	GGACCACAGCAGGAAGAGAC		
	CCL27	Chemokine ligand 27	NM_006664	F	AGCACTGCCTGCTGTACTCA		254
				R	TTCAGCCCATTTTCCTTAGC		
	LOR	Loricrin	NM_000427. 2	F	CTCCTCACTCACCCTTCCTG		114
				R	CCAGAGGTCTTCACGCAGTC		
	KRT9	Keratin 9	NM_000226.3	F	GCTCCTGGCAAAGATCTCAC		155
				R	GACTGCACCTCCTGACCACT		
	CDSN	Corneodesmosin	NM_001264.4	F	CTTGAGCTGCCATCAGTCAG		196
				R	TCGTTAGGGGAGGTGATACG		

Menstrual blood	**MMP7**	Matrix metalloproteinase 7	NM_002423	F	CATGAGTGAGCTACAGTGGGAACAGGC	6-FAM	161
				R	CTATGACGCGGGAGTTTAACATTCCAG		
	MMP10	Matrix metalloproteinase 10	NM_002425	F	ACAGGGAAGCTAGACACTGA		230
				R	CTGGAGAATGTGAGTGGAGT		
	MMP11	Matrix metalloproteinase 11	NM_005940	F	GGTGCCCTCTGAGATCGAC		92
				R	TCACAGGGTCAAACTTCCAGT		
	MSX1	Mshhomeobox 1	NM_002448	F	CCCCGTGGATGCAGAGCCCCCG		96
				R	GCTTACGGTTCGTCTTGTGTTTGCGGAG		
	LEFTY2	Left-right determination factor 2	NM_003240	F	GCCCACGTGAGGGCCCAGTATGTAGT		130
				R	GGTGTGTGCTGGCCTCCGACGC		
	SFRP4	Secreted frizzled-related protein 4	NM_003014	F	GCGACGAGCTGCCTGTCTATGACC		136
				R	CAGTCAACATCAAGAGGCCTTTCCTGTAC		
Vaginal secretion	**MYOZ1**	Myozenin 1	NM_021245	F	GGGTTGGTGAGACAGGATCA	6-FAM	81
				R	TCCCATGGGGAAATATAGGT		
	CYP2B7P1	CytP450, family 2B, polypeptide 7 pseudogene 1	NR_001278	F	TCCTTTCTGAGGTTCCGAGA		198
				R	TTTCCATTGGCAAAGAGCAT		
	MUC4	Mucin 4	NM_018406	F	GGACCACATTTTATCAGGAA		235
				R	TAGAGAAACAGGGCATAGGA		
	HBD1	Human beta-defensin 1	NM_005218	F	CCTGGGTGTTGCCTGCCAGTCGC		200
				R	CAGGTGCCTTGAATTTTGGT		
	Ljen	*Lactobacillus jensenii*	EU559601	F	AAGTCGAGCGAGCTTGCCTATTGAAAT		171
				R	CGCCTTTTAAACTTCTTTCATGCGAAAGTAGC		
	Lcris	*Lactobacillus crispatus*	FN692037	F	GAGAGCAGGAATGCTAAGAG		292
				R	CCGGATCATTGCTTACTTAC		
	Lgas	*Lactobacillus gasseri*	CP000413	F	ATGATGGAGAGTGCGAGAGC		311
				R	CCGGATCATTGCTTACTTAC		
House-keeping	**B2M**	Beta-2-microglobulin	NM_004048	F	GGCATTCCTGAAGCTGACA	ATTO550	120
				R	AAACCTGAATCTTTGGAGTACG		
	UBC	Ubiquitin C	NM_021009	F	GGGTCGCAGTTCTTGTTTGT		186
				R	TCCAGCAAAGATCAGCCTCT		
	UCE	Ubiquitin-conjugating enzyme E2D 2	NM_003339.2	F	AATGATCTGGCACGGGACC		241
				R	ATCGTAGAATATCAAGACAAATGCTGC		

Table 4-3. Primer sequences and amplicon sizes of the 20 mRNA markers in TissueID system (EuroForGen mRNA profiling exercise)

Body fluid	Gene		Primer sequence (5'→ 3')	Dye	Size (bp)
Blood	HBB	F	GCACGTGGATCCTGAGAAC	6-FAM	61
		R	ATGGGCCAGCACACAGAC		
	CD93	F	ACCAGTACAGTCCGACAC	NED	151
		R	TTGCTAAGATTCCAGTCCAG		
	AMICA1	F	TCTCCTGCTCCAAGATGTG	PET	136
		R	GACCATGAGCTCTTTGGG		
Mucosa	KRT4	F	AAAGTCCGGACGGAAGAG	6-FAM	81
		R	TAAGAACTGCACCTTGTCG		
Saliva	HTN3	F	GCAAAGAGACATCATGGGTA	VIC	134
		R	GCCAGTCAAACCTCCATAATC		
	STATH	F	TTTGCCTTCATCTTGGCTCT	6-FAM	93
		R	CCCATAACCGAATCTTCCAA		
Semen	SEMG1	F	GGAAGATGACAGTGATCGT	6-FAM	121
		R	CAACTGACACCTTGATATTGG		
	PRM1	F	AGACAAAGAAGTCGCAGAC	NED	91
		R	TACATCGCGGTCTGTACC		
Vaginal secretion	HBD1	F	GAAATCCTGGGTGTTGCC	6-FAM	101
		R	AAAGTTACCACCTGAGGCC		
	MUC4	F	CTGCTACAATCAAGGCCA	6-FAM	141
		R	AAGGGAAGTTCTAGGTTGAC		
	CYP2B7P1	F	AGTCTACCAGGGATATGGCATG	VIC	146
		R	CTATCAGACACTGAGCCTCGTCC		
Menstrual blood	MMP7	F	GAACAGGCTCAGGACTATCTC	VIC	126
		R	TAACATTCCAGTTATAGGTAGGCC		
	MMP10	F	GCATCTTGCATTCCTTGTGCTGTTG	VIC	107
		R	GGTATTGCTGGGCAAGATCCTTGTT		
	MMP11	F	CAACCGACAGAAGAGGTTCG	NED	76
		R	GAACCGAAGGATCCTGTAGG		
Skin	CDSN	F	CTGGCTGGTCTCCTCCTG	VIC	71
		R	GGGTCCTTACAAGGGTCTGA		
	LCE1C	F	TGTGACCCCGCTCCTGAATCCG	NED	99
		R	CTTGGGAGGGCACTTGGGGGTG		
	LOR	F	CTTTGGGCTCTCCTTCCT	PET	89
		R	AGAGGTCTTCACGCAGTC		
House-keeping	GAPDH	F	GTCCACTGGCGTGTTCACCA	6-FAM	261
		R	GTGGCAGTGATGGCATGGAC		
	18S rRNA	F	CTCAACACGGGAAACCTCAC	PET	110
		R	CGCTCCACCAACTAAGAACG		
	ACTB	F	TGACCCAGATCATGTTTGAG	PET	75
		R	CGTACAGGGATAGCACAG		

Saliva 3plex:

12.5 µl of Multiplex PCR Mastermix (2X) (providing a final concentration of 3 mM $MgCl_2$) (QIAGEN), 2.5 µl of Q Solution (QIAGEN), 2.5 µl of primer mix (HTN3 - 0.2 µM, STATH - 0.4 µM, MUC7 - 0.1 µM), 5.5 µl of H_2O and 2 µl of cDNA were mixed for a total volume of 25 µl. Solutions were denatured at 95 °C for 15 minutes, followed by 35 cycles 94 °C for 30 seconds, 57 °C (+0.2 °C per cycle) for 90 seconds, 72 °C for 60 seconds and a final elongation at 72 °C for 60 minutes.

Semen 5plex:

Each 25 µl reaction consisted of 2.5 µl of Buffer II (10X) (Applied Biosystems), 2.5 µl of dNTPs (10 mM) (Applied Biosystems), 3.25 µl of $MgCl_2$ (25 mM), 0.3 µl of AmpliTaq Gold DNA polymerase (5 U/µl) (Applied Biosystems), 2.5 µl of primer mix (PSA - 0.72 µM, PRM2 - 0.048 µM, SEMG1 - 0.2 µM, TGM4 - 0.6 µM, PRM1 - 0.12 µM), 11.95 µl of H_2O and 2 µl of cDNA. Solutions were denatured at 95 °C for 11 minutes, followed by 35 cycles 94 °C for 20 seconds, 58 °C for 30 seconds, 72 °C for 40 seconds and a final extension at 72 °C for 60 minutes.

Saliva/Semen singleplexes:

Each 25 µl reaction consisted of 2.5 µl of Buffer I (10X) (Applied Biosystems), 1.25 µl of dNTPs (2.5 mM) (Applied Biosystems), 0.25 µl of AmpliTaq Gold DNA polymerase (Applied Biosystems), 2 µl of primer mix (0.8 µM each), 17 µl of H_2O and 2 µl of cDNA. Solutions were denatured at 95 °C for 11 minutes, followed by 35 cycles 94 °C for 20 seconds, 58 °C for 30 seconds, 72 °C for 40 seconds and a final extension at 72 °C for 60 minutes.

MMP/MB/HKG/Vag/Lacto/Skin2 3plexes:

12.5 µl of Multiplex PCR Mastermix (2X) (QIAGEN), 2.5 µl of Q Solution (QIAGEN), 2.5 µl of primer mix (MMP/Vag/Lacto: 2 µM each, MB: MSX1, SFRP4 - 5 µM, LEFTY2 - 2 µM, HKG: B2M, UBC - 2 µM, UCE - 5 µM, Skin2: 8 µM each), 5.5 µl of H_2O and 2 µl of cDNA were mixed for a total volume of 25 µl. Solutions were denatured at 95 °C for 15 minutes, followed by 35 cycles 94 °C for 30 seconds, 55 °C (+0.2 °C per cycle) for 90 seconds, 72 °C for 40 seconds and a final extension at 72 °C for 30 minutes.

HBD1 singleplex:

Each 25 µl reaction consisted of 2.5 µl of Buffer II (10X) (Applied Biosystems), 2.5 µl of dNTPs (10 mM) (Applied Biosystems), 3 µl of $MgCl_2$ (25 mM), 0.4 µl of AmpliTaq Gold DNA polymerase (5 U/µl) (Applied Biosystems), 2 µl of primer mix (20 µM), 9.6 µl of H_2O and 5 µl of cDNA. Solutions were denatured at 95 °C for 11 minutes, followed by 35 cycles 94 °C for 20 seconds, 55 °C for 60 seconds, 72 °C for 45 seconds and a final elongation at 72 °C for 30 minutes.

Skin1 5plex:

12.5 µl of Multiplex PCR Mastermix (2X) (QIAGEN), 2.5 µl of Q Solution (QIAGEN), 2.5 µl of primer mix (LCE1C - 1 µM, LCE1D, LCE2D, IL1F7, CCL27 - 2 µM), 5.5 µl of H_2O and 2 µl of cDNA were mixed for a total volume of 25 µl. Solutions were denatured at 95 °C for 15 minutes, followed by 35 cycles 94 °C for 30 seconds, 58 °C for 90 seconds, 72 °C for 45 seconds and a final extension at 72 °C for 30 minutes.

TissueID 20plex:

12.5 µl of Multiplex PCR Mastermix (2X) (QIAGEN), 5 µl of primer mix, X µl of cDNA and (7.5-X) µl of H_2O were mixed for a total volume of 25 µl, where X is the cDNA input calculated with the serial cDNA input approach. The primer concentrations were as follows: HBB - 0.035 µM, CD93 - 0.25 µM, AMICA1 - 0.15 µM, KRT4, HTN3, GAPDH, ACTB - 0.2 µM, STATH, PRM1 - 0.3 µM, SEMG1, HBD1, MUC4, MMP7 - 0.8 µM, MMP11 - 0.4 µM, CYP2B7P1, MMP10 - 1.6 µM, LCE1C - 0.02 µM, 18S rRNA - 0.025 µM and CDSN, LOR - 0.6 µM. Solutions were denatured at 95 °C for 15 minutes, followed by 33 cycles 94 °C for 20 seconds, 60°C for 30 seconds, 72 °C for 40 seconds and a final extension at 60°C for 45 minutes.

4.2.3 The 'x=n/2' scoring system

As mentioned in section 3.1.3.6, Lindenbergh *et al* proposed a scoring system that could be implemented when interpreting mRNA profiles in casework (Lindenbergh *et al.*, 2013). A standard mRNA results table was designed so that the results obtained from the mRNA replicates could be presented in categorised evaluations for all cell types represented in the 20plex. Five scoring categories were used including the

obvious “observed” and “not observed” but also the supplementary ones “observed and fits with”, “sporadically observed and fits with” and “sporadically observed, no reliable statement possible” (Lindenbergh *et al.*, 2013) [Table 4-4]. Firstly, a set of informative mRNA profiles are generated by the serial input approach followed by replicate analyses of the most informative input. The signals of housekeeping genes can be very useful as they could determine whether a profile is informative. Even though housekeeping genes are expressed across most tissues, their expression levels vary resulting in distinct profiling characteristics. Ideally, four mRNA profiles are obtained from the same sample. Then, the frequency of a signal (of acceptable peak morphology and above detection threshold) (x) relative to the times that a signal could have occurred (n) can be determined for each cell type. The ‘n’ changes according to how many mRNA replicates and number of cell type-specific markers are available.

Absence of peaks is translated to the category “not observed”, while observation of at least half of potential cell type signals for a particular body fluid or tissue results in the term “observed”. An observed cell type is categorised as “observed and fits with” when it is expected to be co-expressed with another cell type. For example, detection of blood markers in a menstrual blood stain is very common; however, extra caution is required so that a ‘true’ presence of blood is not masked. When cell type-specific signals are present in less than half of all possible positions, the term “sporadically observed” is used. Similarly, the categories “sporadically observed and fit with” and “sporadically observed, no reliable statement possible” are employed to assess potential co-expression issues or marker drop-in.

Table 4-4. Scoring and interpretation table

*Scoring:	Interpretation
$x \geq n/2$	Observed
	Observed and fits with
$x = 0$	Not observed
$x \leq n/2$	Sporadically observed, no reliable statement possible
	Sporadically observed and fits with

4.3 Results

Dilution series and body fluid stains were analysed as proposed resulting in either full or partial mRNA profiles. All RT- and PCR negative controls showed no peaks confirming that there was no RNA contamination throughout analysis. As a general remark, certain amplicons were detected as being 1-2 bp longer than expected; although this was suspected to be due to instrumental settings. The results of this study will also be discussed in comparison with the ones obtained by other laboratories (Haas *et al.*, 2014; Haas *et al.*, 2013a; Haas *et al.*, 2015; van den Berge *et al.*, 2014).

4.3.1 EDNAP mRNA profiling exercises

4.3.1.1 Sensitivity of proposed PCR systems

Semen 5plex and SEMG1/TGM4 singleplexes

Using the semen 5plex PCR, as little as 0.5 µl of semen stains were successfully identified with all five markers being amplified [Figure 4-1a]. However, significant variation in expression levels was observed; SEMG1 and TGM4 mRNAs produced the highest peaks. Although other laboratories were able to detect PRM1 in even smaller stains (down to 0.05 µl), in this study no further peaks for any marker were detected when using the pentaplex. These findings are also supported by previously published sensitivity studies (Haas *et al.*, 2009b). Nevertheless, when using the singleplex reactions, the sensitivity of the mRNA markers increased and stains made of only 0.05 µl of semen could be identified using the PRM2 mRNA [Figure 4-1b]. Interestingly, a significant difference between the multiplex and singleplex reactions was observed, since PRM2 was the least sensitive as part of the 5plex but demonstrated the highest sensitivity when amplified alone. A possible explanation could be the different PCR conditions as in the singleplex there is no longer competition between mRNA molecules regarding the PCR components (e.g. dNTPs).

Saliva 3plex and HTN3/MUC7 singleplexes

It should be noted that in all samples a larger amplicon (152 bp) was also detected that corresponds to a histatin isoform (HTN1), which has 95% sequence identity to marker HTN3 [Figure 4-2a]. This has been previously reported in literature and does not interfere with data analysis (Haas *et al.*, 2009a; Juusola & Ballantyne, 2005).

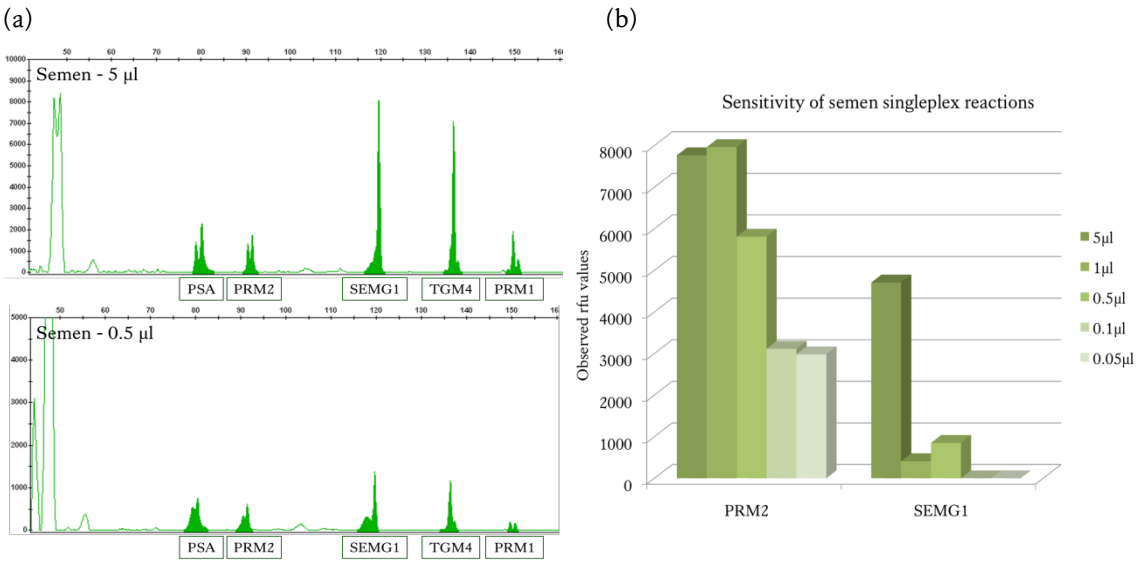


Figure 4-1. Sensitivity of the semen mRNA markers

(a) Electropherograms of two semen stains (5 and 0.5 µl semen respectively) analysed with the semen pentaplex; all five markers were detected, although some peaks were saturated resulting in split peaks, (b) Peak heights obtained by analysing the semen dilution series (5, 1, 0.5, 0.1, 0.05 µl semen) with both singleplexes (PRM2 and SEMG1); as shown, PRM2 was detected in as little as 0.05 µl semen stain (2,979 rfu).

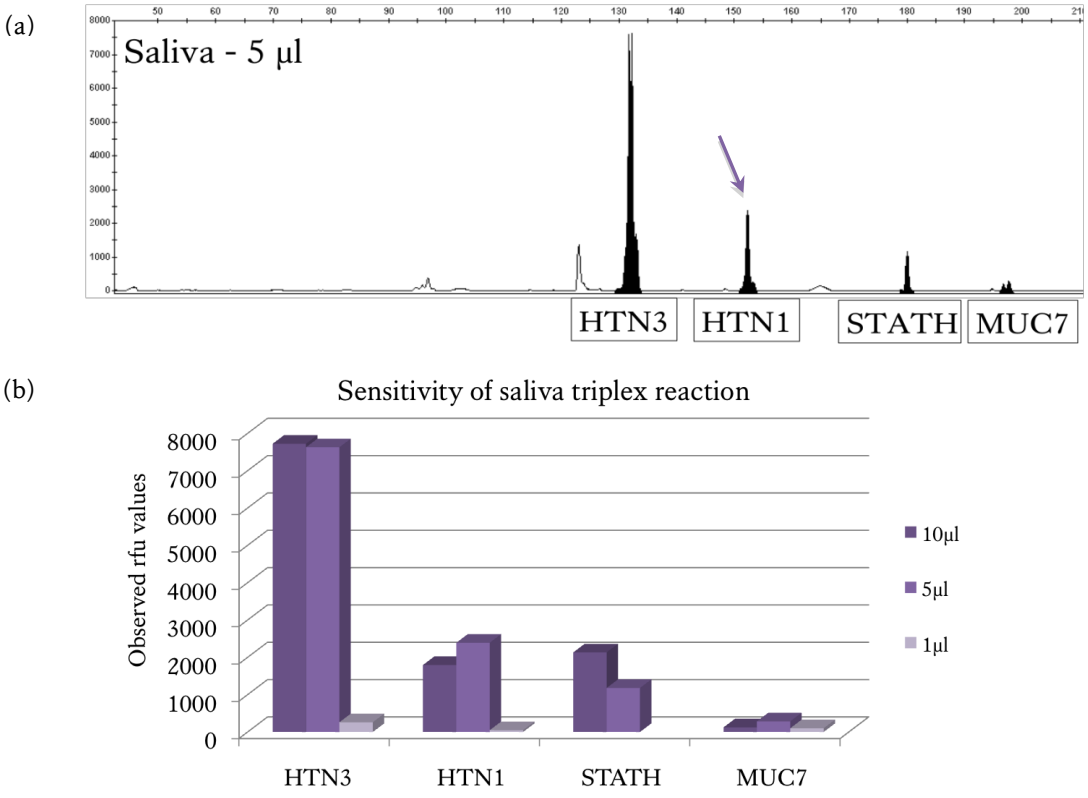


Figure 4-2. Sensitivity of the saliva mRNA markers

(a) Electropherogram of a 5 µl saliva stain analysed with the saliva triplex; as shown, all three markers were detected (plus the isoform HTN1, indicated with the arrow), (b) Obtained peak heights of the saliva dilution series (10, 5, 1 µl saliva) analysed with the saliva triplex; significant variation (10-fold) in peak heights was observed.

Saliva stains were successfully identified down to 1 μl of saliva [Figure 3-3b], whereas other laboratories achieved detection down to 0.05 μl of saliva. In contrast with other groups, MUC7 demonstrated the least sensitivity since no peaks were detected using the singleplex assay.

Menstrual blood MMP/MB 3plexes

For the identification of menstrual blood, two 3plexes were tested; however, a significant variation in specificity and sensitivity was observed amongst the mRNA markers. The metalloproteinase MMP 3plex proved to be the most sensitive, since successful identification of menstrual blood was achieved by detecting MMP7 using just 1/64 of a stained cotton swab [Figure 4-3b]. Additionally, all three metalloproteinases (MMP7, MMP10 and MMP11) were successfully amplified (250 rfu, 334 rfu and 511 rfu respectively) down to 1/32 of a cotton swab. Comparing the results obtained from other laboratories, 14 out of 20 were able to detect MMP10 when extracting just 1/64 of a menstrual blood swab. On the other hand, the MB 3plex did not show similar sensitivity since only MSX1 was detected down to 1/16 of a stained cotton swab. Remarkably, the marker SFRP4 was not detected in any of the dilution series samples indicating its lack of sensitivity for menstrual blood identification. Less than 15% of the laboratories identified this marker in their samples.

It is believed that the markers included in the MB 3plex change their expression depending on the day of the menstrual cycle indicating that metalloproteinases might be the most suitable markers for the identification of menstrual blood. However, possible intra- and inter-individual variation in gene expression was evident and should be taken into account since it is known that menstrual blood flow can vary significantly (Fraser *et al.*, 1985). Also, since the dilution series were prepared in a more 'qualitative' way rather than quantitative (portions of swab rather than exact volumes), variations in the amount of recovered menstrual blood material cannot be excluded either.

Vaginal Vag/Lacto 3plexes and HBD1 singleplex

Two 3plexes and one singleplex were tested for vaginal secretion and the results were promising. All selected mRNA markers apart from Lgas included in the Lacto 3plex demonstrated high sensitivity and were successfully identified in the smallest sample

size (1/64 of a vaginal fluid swab) [Figure 4-4b]. Lgas was only sporadically observed by other laboratories as well, possibly due to the limited co-presence of different *Lactobacilli* species in the sample used for the dilution series and not as a result of low detection sensitivity. As shown in Figure 4-4, Ljen seemed to be the most sensitive and abundant mRNA marker since it was constantly detected resulting in very high peaks. Lastly, it should be noted that the sample containing 1/32 of a vaginal swab resulted in lower peaks for most of the markers compared to the 1/64 sample, either due to particular experimental circumstances or potentially because of an 'uneven' distribution of fluid material over the swab when dilution series were prepared.

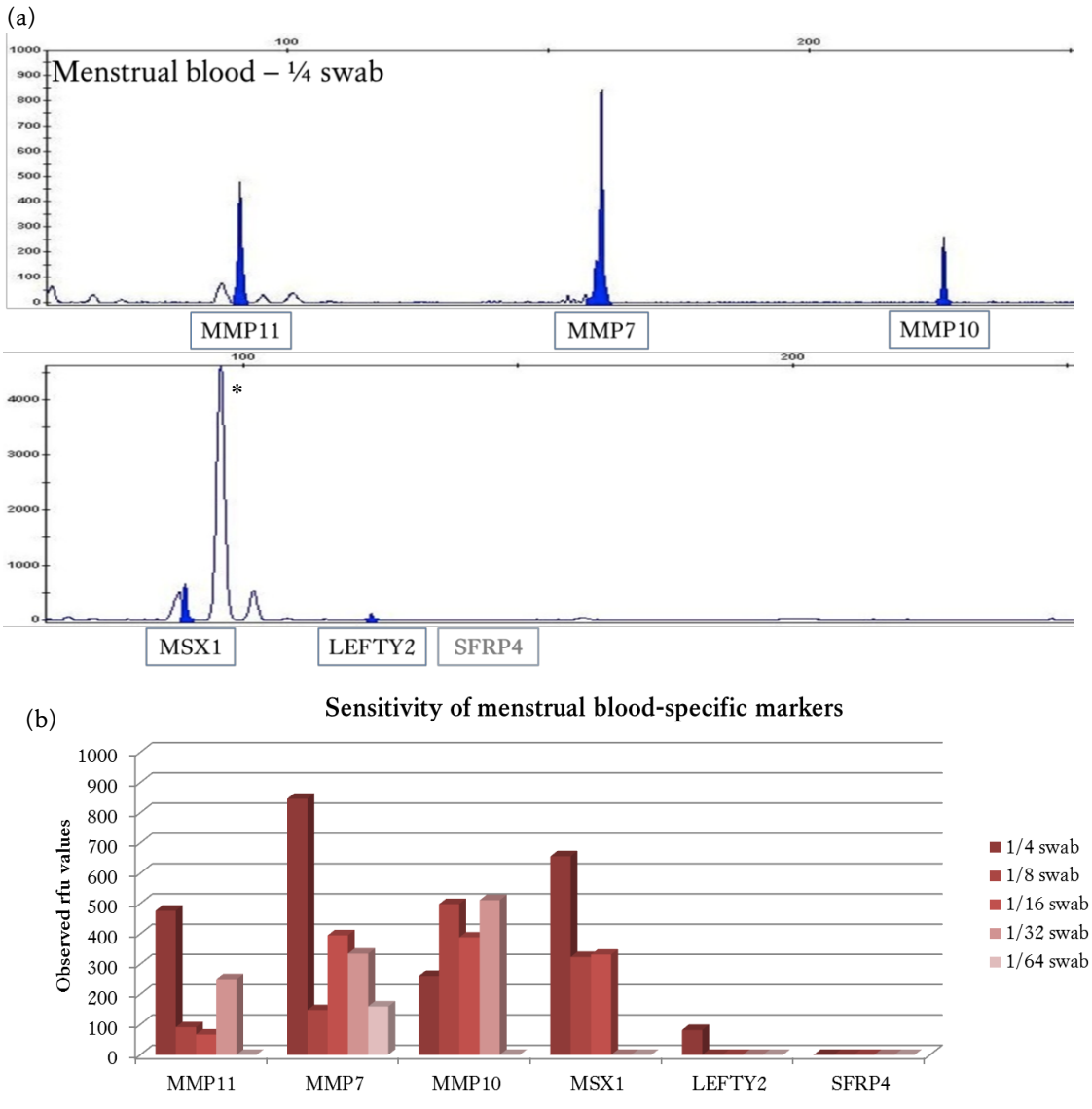


Figure 4-3. Sensitivity of the menstrual blood mRNA markers

(a) Electropherogram of a ¼ menstrual blood swab analysed by both menstrual blood triplexes; as shown, all three markers were detected for MMP 3plex, and only two (unsuccessful amplification for SFRP4) for the MB 3plex (the asterisk (*) indicates a non-specific peak due to dye blob, present in both samples and negative control), (b) Obtained peak heights of the menstrual blood dilution series (1/4, 1/8, 1/16, 1/32, 1/64 of a cotton swab) analysed with both multiplexes; as demonstrated, significant variation in sensitivity was observed.

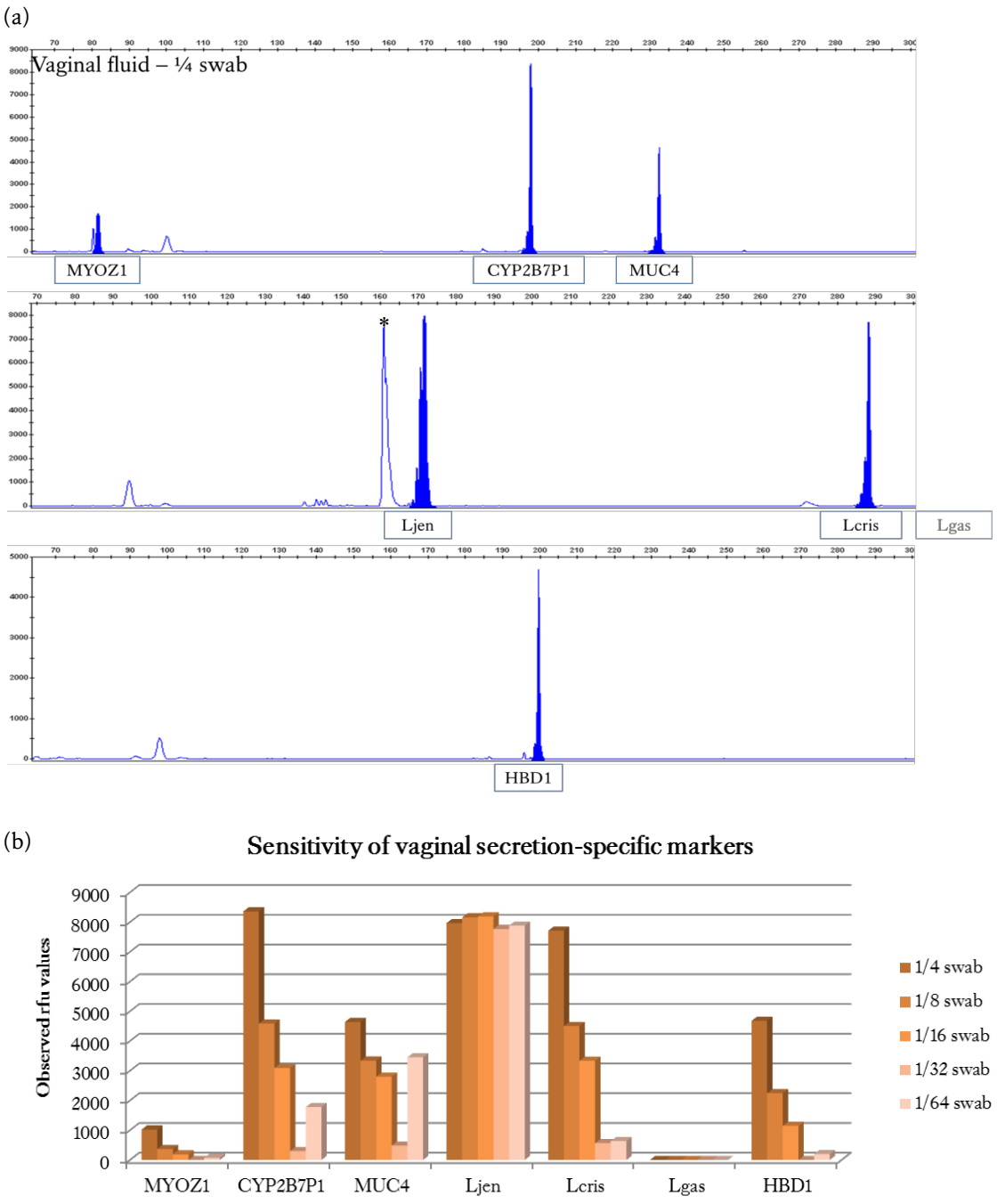


Figure 4-4. Sensitivity of the vaginal fluid mRNA markers

(a) Electropherogram of a ¼ vaginal fluid swab analysed by both vaginal fluid triplexes plus the HBD1 singleplex; as shown, all markers were detected apart from the Lgas in Lacto 3plex (the asterisk (*) indicates a non-specific peak, more likely resulting from the amplification of an isoform, present in both the samples and negative control), (b) Obtained peak heights of the vaginal fluid dilution series (1/4, 1/8, 1/16, 1/32, 1/64 of a cotton swab) analysed with all assays; as expected, some markers seemed to be more abundant and sensitive than others.

Skin Skin1 5plex/Skin2 3plex

Two different multiplexes – a 5plex and a 3plex – were tested for their ability to successfully identify touch samples. It should be noted that the amplification of the LCE1C marker resulted in a split peak in most samples [Figure 4-5]. Also, since skin

samples are generally considered of low quality and quantity, the post-PCR purification step was omitted resulting in a number of non-specific peaks. It is believed that these peaks cannot interfere and affect interpretation since they are detected at different lengths and are observed in both samples and negative control. As shown in Figure 4-5, five out of the total eight mRNA markers were successfully detected even at the smallest amount of sample (0.8 ng skin RNA), while CCL27 and CDSN were still amplified at 3 ng RNA sample. Interestingly, KRT9 was only detected at 200 ng giving a relatively low peak (101 rfu), therefore showing the lowest sensitivity between the proposed markers. Undoubtedly, LCE1C and LOR were found to be the most sensitive skin markers and this finding was consistent across all laboratories.

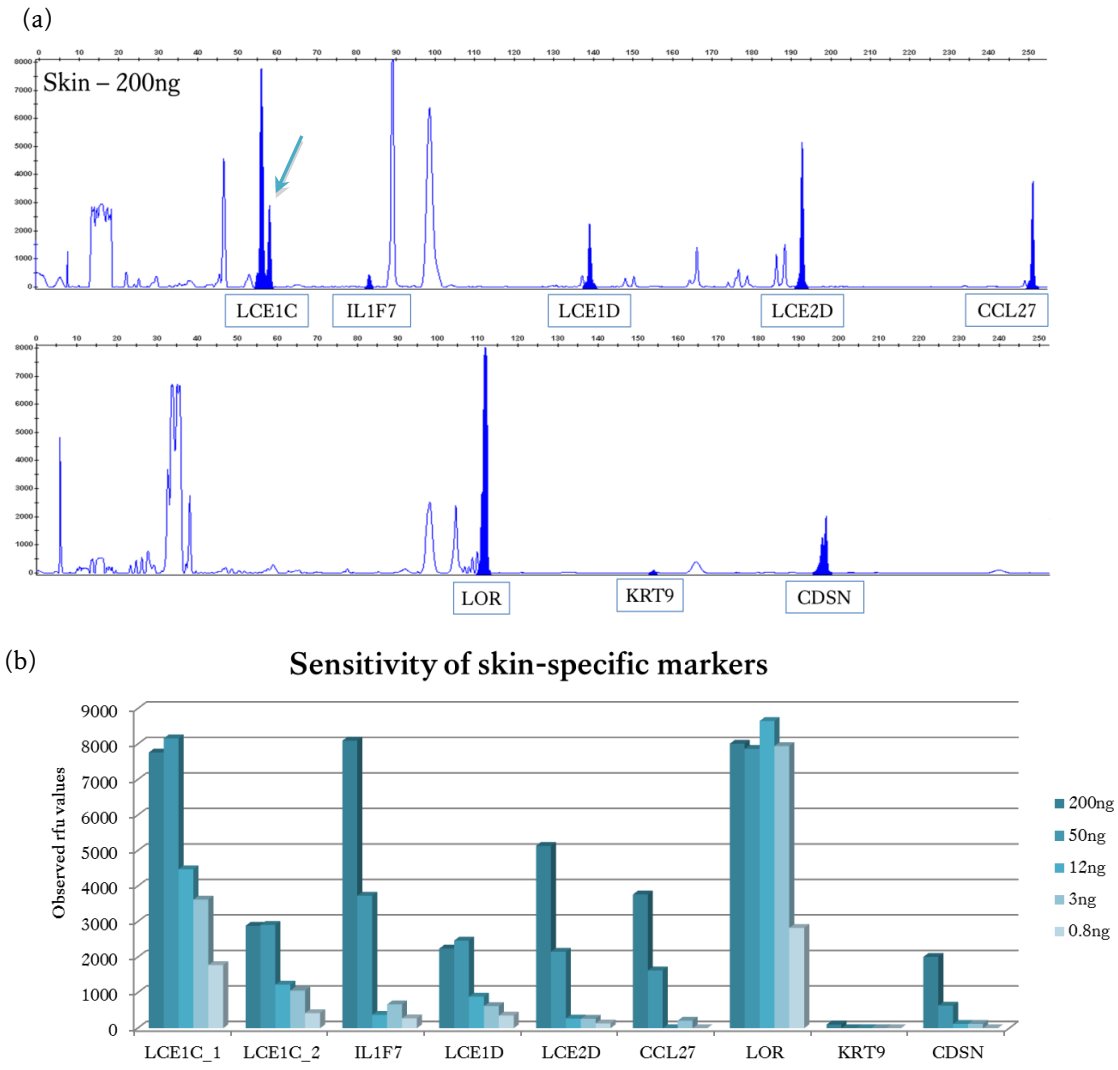


Figure 4-5. Sensitivity of skin mRNA markers

(a) Electropherogram of the sample containing 200 ng of skin RNA analysed by both multiplexes – Skin1 pentaplex and Skin2 triplex; as shown, all markers were amplified but there was significant variation in peak heights (all non-coloured peaks are non-specific) (b) Obtained peak heights of the skin dilution series (200, 50, 12, 3 and 0.8 ng skin RNA) analysed with both assays; most markers (apart from KRT9) showed great sensitivity since most of them were successfully detected down to 0.8 ng.

Housekeeping HKG 3plex

The definitive identification of the presence of cellular material was successfully accomplished with the use of a HKG 3plex. To assess the sensitivity of the assay, decreasing amounts of blood, semen, saliva, menstrual blood and skin were tested. As expected, peaks for these housekeeping genes were obtained with as low as 0.01 μl of blood, 0.2 μl of semen, 1 μl of saliva, 1/64 of a menstrual blood and 0.8 ng of skin RNA. B2M seemed to be the most sensitive in most tissues, while UCE was the least sensitive.

4.3.2 Marker specificity and inter-individual variation

The applicability of each proposed multiplex was investigated by analysing a specific blind test of body fluid stains [Table 4-1], which could contain the target tissue, non-target tissues or non-human cells. Samples also included mixtures, low quality and aged body fluid stains.

Ten body fluid stains (set 1) were analysed using the semen 5plex and saliva 3plex PCR systems. In general, no false positives were observed and only one stain resulted in no peaks [Table 4-5a]. For stain 1 (5 μl of saliva on paper) none of the saliva markers were detected; indeed, only four out of the 19 laboratories managed to predict that there was saliva on the paper. This stain appeared to be challenging possibly because of the presence of inhibiting substances in the paper. The effect of the paper dye in the PCR efficiency should be emphasized, as it has been previously noted to act as an inhibitor (data not shown). Stain 2 was semen from an azoospermic male and therefore, only the seminal markers (and not PRM1/PRM2) were expected. Also, no cross-reactions were detected in the vaginal swab (stain 3), highlighting saliva and semen markers' specificity. Stain 4 (5 μl of saliva on glass slide) was successfully identified; however, STATH mRNA was undetectable. Since it had been detected in 5 μl of saliva when analysing the dilution series, the absence of the STATH peak could be due to low expression levels in the donor's saliva.

Table 4-5. mRNA profiling of 44 body fluid stains using the proposed EDNAP multiplexes

Four sets of stains were analysed with the corresponding multiplexes and the peak heights (rfu) for each marker are presented. Coloured squares (purple, green, red, orange and blue) represent the markers that were expected to be detected using each multiplex (Saliva, Semen, MMP/MB, Vag/Lacto/HBD1 and Skin1/Skin2), while the ones highlighted in grey were not expected to give any signal. The peak heights of non-specific signals are emphasised in red.

(a)	Stains	Saliva 3plex (rfu)			Semen 5plex (rfu)				
		HTN3	STATH	MUC7	PSA	PRM2	SEMG1	TGM4	PRM1
1	5µl saliva on paper	-	-	-	-	-	-	-	-
2	3µl azoospermic semen on pad	-	-	-	770	-	2429	1650	-
3	1/4 vaginal swab	-	-	-	-	-	-	-	-
4	5µl saliva on glass slide	403	-	72	-	-	-	-	-
5	2µl semen on toilet paper	-	-	-	48	79	-	-	215
6	1µl semen/5µl saliva on swab	7440	695	171	369	122	2358	-	111
7	1/4 buccal swab (cat)	-	-	-	-	-	-	-	-
8	1µl blood/2µl saliva on swab	238	-	178	-	-	-	-	-
9	licked plastic spoon	4394	-	155	-	-	-	-	-
10	3µl semen inside latex glove	-	-	-	-	53	61	-	169

(b)	Stains	MMP 3plex (rfu)			MB 3plex (rfu)		
		MMP11	MMP7	MMP10	MSX1	LEFTY2	SFRP4
1	Menstrual blood on sanitary towel (fresh)	2445	390	4244	51	1231	-
2	¼ cotton swab with EDTA-blood (fresh)	136	-	-	-	-	-
3	¼ menstrual blood swab (fresh)	9408	9152	9298	7497	8405	8541
4	Menstrual blood on sanitary towel (5 years old)	8174	-	3761	55	1443	-
5	¼ vaginal swab (fresh)	-	3207	-	101	-	-
6	¼ menstrual blood swab (5 years old, day 1-4)	132	198	217	50	55	-
7	¼ menstrual blood swab (5 years old, day 1/day 4)	963	102	1804	116	92	57
8	¼ menstrual blood swab (fresh)	40	-	234	-	-	59
9	Menstrual blood swab (day 3)	8052	3573	4896	461	103	-
10	Suspected saliva stain (unknown volume)	-	-	-	-	-	-

(c)	Stains	Vag 3plex (rfu)			Lacto 3plex (rfu)			Singleplex
		MYOZ1	CYP2B7P1	MUC4	Ljen	Lcris	Lgas	HBD1
1	¼ vaginal swab (2 years old)	979	8333	4567	462	8131	-	8118
2	5x5 mm from white worn underpants (fresh)	- (36)	1390	777	8232	7626	-	-
3	½ swab with urine (fresh)	-	-	-	8325	-	91	-
4	½ vaginal swab, pregnant (fresh)	1570	8259	8303	788	-	7929	8014
5	¼ vaginal swab (5 years old)	364	8076	2767	5792	7765	-	1259
6	½ vaginal swab (fresh)	971	7820	7577	8010	-	7896	4679
7	½ buccal swab (fresh)	-	-	154	8702	-	-	443
8	5x5 mm from sanitary towel (fresh)	-	-	-	7735	-	-	497
9	Vaginal swab (day 16)	1063	8311	8282	8055	3865	-	7855
10	½ menstrual blood (day 4)	-	7808	7818	8011	5533	-	3199
11	Skin swab from palm	-	-	-	2175	-	-	-
12	5µl saliva stain	-	-	-	115	-	-	-

(d)	Stains	Skin1 5plex (rfu)					Skin2 3plex (rfu)		
		LCE1C	IL1F7	LCE1D	LCE2D	CCL27	LOR	KRT9	CDSN
1	Small swab from palm	6160/3780	267	1285	512	-	7866	-	-
2	Hand print on paper, glossy side	1140/259	-	-	-	-	-	-	-
3	Key from computer keyboard	-	-	-	-	-	-	-	-
4	Fingerprint on a glass slide	-	-	-	-	-	-	-	-
5	Small swab with urine	-	-	-	-	-	-	-	-
6	Swab from palm/1 µl of blood	6183/1209	277	537	136	-	7449	311	57
7	Small swab with saliva	185/223	-	-	-	-	136	-	-
8	Scraped skin from the back of the hand	5858/2693	4207	1274	5314	822	7836	155	2178
9	Skin swab from palm (fresh)	7408/2339	467	1859	359	-	8120	-	143
10	5µl saliva stain	-	-	-	-	-	209	-	-
11	Vaginal swab (day 13)	7684/3646	136	1361	825	-	7907	-	1075
12	Hair (including the root)	419/-	-	- (40)	-	-	-	-	-

Furthermore, three out of the five semen markers were successfully detected in stain 5 (2 μl of semen on toilet paper), namely PSA, PRM1 and PRM2; however, the peak heights were considered quite low. On the other hand, when using the PRM2 singleplex a 2,609 rfu peak was obtained. Stain 6 was successfully identified as a saliva/semen mixture; interestingly, only half of the laboratories detected TGM4. The buccal swab from a cat (stain 7) gave no peaks when tested with both multiplexes, indicating that the primers are also human-specific. Similarly to stain 2 results, stain 8 (blood/saliva mixture) and 9 (used spoon) were correctly identified, however, STATH was again not detected. Lastly, stain 10 (3 μl semen inside latex glove) was differentiated by detecting three out of five markers and no saliva markers. Similar results were noticed in stain 5 in terms of PRM2 singleplex; a 4,771 rfu peak was obtained by the singleplex, whereas the semen 5plex gave a very low 53 rfu peak.

The next set of ten stains (eight EDNAP and two in-house samples) were analysed using the menstrual blood MMP and MB PCR systems. In general, no false negatives were observed; however, there were cases when menstrual blood markers were detected in non-target tissues [Table 4-5b]. More specifically, markers MMP11 and MSX1 showed very low peaks in stain 12 (¼ of a cotton swab with fresh EDTA-blood) and stain 15 (¼ of a fresh vaginal swab) respectively, which is believed to be due to natural variation in gene expression amongst individuals. On the other hand, MMP7 resulted in a high peak for the vaginal fluid that could be explained by the presence of traces of menstrual blood in that particular sample.

Menstrual blood samples demonstrated a significant variation in peak heights, which could be explained by potential inter-individual differences in gene expression or changes during the menstrual cycle. In general, the MMP 3plex proved to be more robust than the MB 3plex since it resulted in less marker drop-outs and higher peaks. It was observed that by day 3 and 4 of menstruation, some these markers (MSX1, LEFTU2, SFRP4) became undetectable; these findings were also supported by the other laboratories. In more detail, stain 13 (1/4 fresh menstrual blood swab) was the only one that showed high signals for both triplexes with an average peak height of 8,717 rfu and strong amplification of the SFRP4 marker. Stain 11 (fresh menstrual blood on sanitary towel), stain 17 (1/4 of 5-years-old menstrual blood swab) and an in-house stain (fresh menstrual blood swab – day 3) highly expressed all

metalloproteinases, however the expression for the markers included in the MB assay had often a 10-fold decrease in peak height. Storage time (up to 5 years) did not seem to be an essential factor for the identification since stain 14 (1/4 of 5-years-old menstrual blood swab) resulted in the amplification of four out of six markers, while in stain 18 (1/4 of fresh menstrual blood swab) only three markers were detected (average peak height 111 rfu).

Similar findings were obtained regarding the identification of vaginal material. Twelve stains including eight EDNAP and four in-house samples were analysed [Table 4-5c]. Although there were no false negative results, the specificity of certain markers (Ljen, MUC4 and HBD1) was questioned due to observed cross-reactivity with other tissues. Also, significant variation in gene expression was observed by all laboratories. The Vag 3plex demonstrated great robustness; nevertheless, all three markers were not detected in stain 26 (5x5 mm from sanitary towel). For the Lacto 3plex, Ljen showed the highest peaks but it is considered an unsuitable vaginal marker, as traces of these bacteria were also observed in urine, menstrual blood, skin and saliva. Similarly to the sensitivity analysis results, Lgas was only found in fresh samples (stains 22 and 24).

Occurrences of cross-reactivity cannot be disregarded; however, the non-specific signal intensities were slightly reduced compared to those in vaginal secretion samples. For example, HBD1 had an average peak height of 5,070 rfu in vaginal fluid and gave a positive signal for buccal epithelium of 443 rfu. Since saliva and urine contain cells of epithelial origin, some cross-reactivity was expected. Also, given that healthy menstrual blood usually contains traces of vaginal fluid, the amplification of five out of seven vaginal mRNA markers was predictable making the identification of vaginal fluid very challenging.

Lastly, regarding the differentiation of touch evidence, twelve stains (eight EDNAP and four in-house samples) were tested using the two proposed skin multiplexes [Table 4-5d]. For the high input samples such as stain 27 (small swab from palm), stain 32 (swab from palm/1 µl blood), stain 34 (scraped skin from the back of the hand) and the in-house stain (skin swab from palm), most skin markers were detected. However, contact traces such as stain 28 (hand print on paper), stain 29 (key from computer keyboard) and stain 30 (fingerprint on a glass slide) proved challenging and difficult to identify (with only LCE1C being detected in stain 2). The presence of a low amount

of biological material for these samples was also confirmed by the HKG 3plex, since only one housekeeping marker (B2M) was amplified in stain 28.

The specificity of the proposed skin-specific markers was tested by analysing urine (stain 31), saliva (stain 33 and in-house sample) as well as vaginal fluid (in-house sample). Whereas urine gave no peaks, LOR and LCE1C were detected in saliva (140 and 220 rfu respectively). Also, skin was strongly detected in the vaginal sample, possibly due to the way that vaginal swabs were collected. The possible presence of skin in saliva or vaginal fluid, or the presence of common cell types in these tissues cannot be excluded; that way the observed different levels of gene expression could be explained by distinct sets of transcription regulators. Finally, the hair sample gave a positive signal for one of the most sensitive skin markers (LCE1C) but this was expected since the hair was pulled hence it could contain skin cells from the scalp.

4.3.3 DNA results

Although not compulsory, DNA extracts were also analysed to assess that the proposed extraction method was efficient not only for differentiating the tissue of origin but also for identifying the donor of the stain through standard DNA profiling. Following DNA/RNA extraction, DNA extracts of all stains and dilution series were eluted in 80 μl of EB buffer and DNA concentrations were quantified using the Quantifiler Human DNA quantification kit (Life Technologies). Wherever possible, DNA samples were diluted down to 1 ng/μl prior to PCR. STR profiles were then obtained using the Powerplex ESI 16 system (Promega).

4.3.3.1 Dilution series

Figure 4-6 shows the correlation of starting body fluid volume (μl/part of swab) with the total DNA recovery (ng) for semen, saliva, menstrual blood and vaginal fluid obtained by the analysed dilution series. As expected, the same volume of semen resulted in higher DNA yield compared to saliva (~6-fold increase); for instance, 1 μl of saliva yielded 4.1 ng of DNA (0.05 ng/μl), while 1 μl of semen produced a total of 23.2 ng of DNA (0.29 ng/μl). Similarly, the same part of a menstrual blood swab yielded more DNA than the corresponding one stained with vaginal fluid (~6-fold increase); for example, 1/8 of menstrual blood swab yielded a total of 2.42 μg of DNA (30.27 ng/μl), whereas 1/8 of a vaginal swab resulted in 0.44 μg of DNA (5.45 ng/μl).

As illustrated, the relationship between sample volume and DNA yield was linear for all body fluids (R^2=0.95-0.99) suggesting that the followed DNA/RNA co-extraction method performed with the same efficiency regardless of the starting cellular material. Full profiles were obtained in the saliva dilution series from 1 μl of starting material and in the semen dilution series from as little as 0.5 μl of semen. Also, 1/64 of both a menstrual or vaginal swab resulted in a full DNA profile.

4.3.3.2 Body fluid stains

As expected, the body fluid stains resulted in a range of DNA concentrations depending on the quantity and quality of the starting biological material. For the stain set 1 (EDNAP stains 1-10, saliva/semen), most resulted in full STR profiles that matched the reference profiles. As expected, stain 7 (cat) and 2 (azoospermic semen) gave no profiles. Regarding stain set 2 (EDNAP stains 11-18, menstrual blood), most resulted in full STR profiles; interestingly, stain 14 (sanitary towel) and 18 (fresh menstrual blood) gave rise to partial profiles. In stain set 3 (EDNAP stains 19-26, vaginal fluid), all stains also resulted in full DNA profiles apart from urine (partial profile). Lastly, stain set 4 (EDNAP stains 27-34, skin) seemed the most challenging one due to the low quantity of touch samples; full profiles were obtained only for stains 32 (1 μl blood), 33 (swab with saliva) and 34 (scraped skin). All touch samples (stains 27-30) yielded between 0.06-0.46 ng of DNA in total, which is known to be insufficient for successful STR analysis.

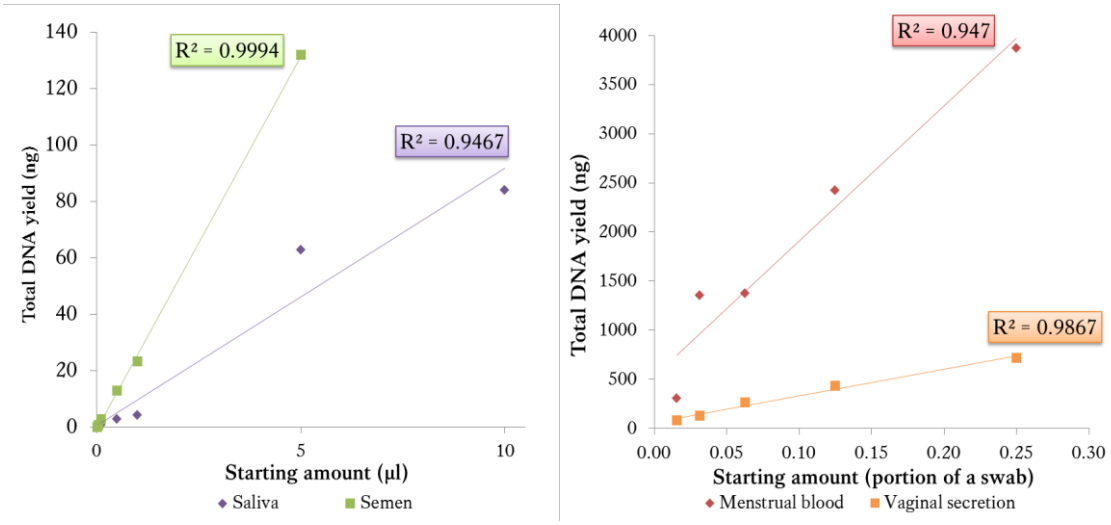


Figure 4-6. Total DNA yield versus starting volume for all body fluids tested

4.3.4 EuroForGen mRNA profiling exercise

4.3.4.1 Sensitivity of multi-tissue 20plex

In order to test the sensitivity of the proposed multi-tissue 20plex, 24 single-source cDNA samples derived from three mRNA inputs (plus a RT- control) for each of the six cell types (blood, semen, saliva, menstrual blood, vaginal secretion and skin) were analysed. Samples were prepared as mentioned in section 4.2.1.4 and only 1 µl of cDNA was used for all PCR reactions. This analysis also helped with assessing potential differences in marker performance within the 20plex. Results are summarised in Table 4-6. Most markers seem to be very sensitive, however a few drop-outs were observed. Although findings could be affected by the use of a single donor for each tissue type, certain conclusions regarding the assay performance could be made. With reference to the housekeeping genes, 18S rRNA appeared to be the most robust and GAPDH appeared the least robust marker; findings that are also supported by the other laboratories.

In general, the 20plex performed better than the previously analysed cell type-specific multiplexes as detection was achieved down to 0.2 µl of extracted RNA. Nevertheless, since RNA solutions were not quantified, conclusions regarding sensitivity should be made with caution. The most interesting and challenging body fluid seemed to be the vaginal fluid, where only CYP2B7P1 was amplified across all three mRNA inputs, while HBD1 and MUC4 were not amplified at all. This could suggest suboptimal performance of these two markers in the 20plex, which has previously been discussed (Lindenbergh *et al.*, 2012). Additionally, non-specific signals were observed for vaginal mucosa: (1) blood (CD93 in particular), which may have been the result of traces of menstrual blood (the only amplified blood marker in menstrual blood samples) and (2) skin (LOR in particular), which could either question the specificity of this skin marker or indicate the 'true' presence of skin cells in the vaginal swab.

Table 4-6. Sensitivity of the TissueID 20plex

Detected markers when analysing cDNAs derived from three inputs of single source RNAs. For each sample, a profile was obtained by using 1 µl of cDNA input. Green boxes indicate successful detection of expected markers, while red indicate non-specific amplification.

RNA input (µl)		Blood			Semen			Saliva			Menstrual blood			Vaginal fluid			Skin		
mRNA markers		1	0.4	0.2	1	0.4	0.2	1	0.4	0.2	1	0.25	0.125	1	0.4	0.2	1	0.4	0.2
Blood	HBB	■	■	☒							☒	☒	☒						
	CD93	■	☒	■							■	■	☒	■		■			
	AMICA1	☒	☒	☒							☒	☒	☒						
Semen	SEMG1				■	■	■												
	PRM1				■	■	■												
Saliva	STATH							■	☒	■									
	HTN3							■	■	☒									
Menstr bl	MMP10										■	■	■						
	MMP7										■	■	☒						
	MMP11										■	☒	☒						
Vag secr	HBD1										■	☒	☒	☒	☒	☒			
	MUC4										■	■	■	☒	☒	☒			
	CYP2B7P1										■	■	■	■	■	■			
Skin	CDSN																■	■	■
	LCE1C																■	☒	■
	LOR													■	■	■	■	☒	■
M	KRT4							■	■	☒	■	■	■	■	■	■			
HKG	ACTB	■	■	■	■	☒	☒	☒	☒	☒	■	■	■	■	■	■	■	■	☒
	18S-rRNA	■	■	■	■	■	■	■	■	☒	■	■	■	■	■	■	■	■	■
	GAPDH	■	■	☒	☒	☒	☒	☒	☒	☒	■	■	■	■	■	☒	☒	☒	☒

■ Expected and detected

☒ Expected but not detected

■ Detected but not expected

4.3.4.2 Impact of cDNA input

The amount of cDNA template added into the PCR could have an impact on the number of markers amplified, hence the number of amplicons detected. To illustrate the effect of cDNA input on RNA profiling results, eight single-source cDNA samples and a mixture were tested using the TissueID 20plex. Two mRNA profiles per sample were generated using a four-fold difference in cDNA input volume (0.5 µl and 2 µl respectively). In theory, a total of 32 cell type-specific peaks (without the mucosa and housekeeping markers) should have been detected for all nine cDNA specimens with each cDNA input. Using 0.5 µl of cDNA, 14 peaks (43.8%) were detected, whereas with the higher volume (2 µl), 24 peaks (75%) were observed. Cell type-specific peaks

detected using both the 0.5 µl and 2 µl input (15 in total) were on average five times higher for the latter, which relatively complies with the four-fold increased cDNA input volume. Marker drop-out was predominantly seen for the vaginal secretion marker HBD1 as well as the blood marker AMICA1, both of which are among the least sensitive markers as shown above. Table 4-7 illustrates the impact of cDNA input in the detection of all 20 markers included in the assay; in conclusion, a four-fold increase in the cDNA amount resulted in an average increase of 38.7% in the number of detected peaks; however, marker drop-in (skin) was observed in the vaginal secretion samples when using higher amount of cDNA.

Furthermore, since some blood, saliva and skin samples used in the previous section (Table 4-6) had been prepared under the same experimental conditions but with using 1 µl of cDNA input, it was thought that the results obtained from 1 and 2 µl of cDNA could be further compared. Interestingly, for these three tissues there were no significant changes in detected markers between the two cDNA inputs; however similar conclusions cannot be made for the rest of the tissues. It is believed that 0.5 µl is a small amount that should be used with caution as cDNA input volume since it could lead to marker drop-outs due to pipetting errors or insufficient template.

Table 4-7. Impact of cDNA input in the detection of all expected peaks for each tissue
Eight single-source samples (two skin samples and an azoospermic semen sample) together with a vaginal fluid/blood mixture were analysed using two cDNA inputs (0.5 µl and 2 µl). The expected and observed peaks were used to demonstrate the impact of cDNA input in the efficiency of the assay.

Sample	cDNA (µl)	Peaks				Efficiency
		Expected	Observed	Missed	Unexpected	
Blood	0.5	6	3	3	0	50.0%
	2	6	5	1	0	83.3%
Saliva	0.5	6	0	6	0	0.0%
	2	6	5	1	0	83.3%
Skin 1	0.5	6	3	3	0	50.0%
	2	6	5	1	0	83.3%
Skin 2	0.5	6	1	5	0	16.7%
	2	6	4	2	0	66.7%
Vaginal secretion	0.5	7	2	5	0	28.6%
	2	7	5	2	2	71.4%
Menstrual blood	0.5	13	10	3	0	76.9%
	2	13	10	3	0	76.9%
Semen	0.5	5	3	2	0	60.0%
	2	5	5	0	0	100.0%
Infertile semen	0.5	4	2	2	0	50.0%
	2	4	3	1	0	75.0%
Vaginal secretion/blood	0.5	10	4	6	1	40.0%
	2	10	8	2	2	80.0%

4.3.4.3 Applicability in forensic casework

As previously shown in section 4.3.2, testing single- or double-source body fluid stains proved to be somewhat difficult. The low yields of biological material from aged or degraded samples, the potential cross-reactivity of certain cell-type specific mRNA markers and the significant inter-individual variation in gene expression levels make the application of mRNA profiling a challenging task. The need for a robust experimental strategy including a suitable scoring system and interpretation guideline is clear.

Usefulness of the 'x=n/2' scoring system

In order to test the utility of the proposed scoring system, four cDNAs coming from mixtures of two or three cell types in balanced (1:1) or unbalanced (up to 1:10) ratios were analysed. All samples were tested using both 0.5 µl and 2 µl of cDNA input; however, the latter was considered as the best option since some marker drop-outs were observed when using 0.5 µl. Afterwards, using 2 µl of each cDNA four informative mRNA replicates were generated and analysed using the scoring system [Table 4-8].

Briefly, cDNA 1 was correctly identified since both blood and saliva were "observed"; however, sporadic peaks for vaginal secretion (3 peaks) and skin (1 peak) were also detected. All three components (blood, saliva and skin) were also correctly determined in cDNA 2; nevertheless, three peaks for menstrual secretion and vaginal mucosa were detected. From these results, it becomes apparent that it is helpful to employ the category "sporadically observed" (generally regarded as not observed) as it can prevent false positive identification. Detecting six out of the twelve possible skin peaks in cDNA 3 (10:1 vaginal fluid:semen) led to a false positive identification of skin; a finding confirmed by all participating laboratories. This seems a frequent event, possibly a consequence of the unintended responses of LOR (and to a lesser extent CDSN) in vaginal secretion. Therefore, it seems beneficial to exclude LOR results from interpretation. Lastly, regarding cDNA 4, the lowest component of the mixture (skin) was missed as it was only "sporadically observed". Notably, the assay allowed for the identification of the azoospermic semen.

Table 4-8. Interpretation of four mixed mock casework samples using the scoring system
Colour coding of interpretation cells is: dark green = correct, i.e. not observed when not present or observed (& fits) when present; light green = sporadically observed (regarded as not observed) when not present; red = incorrect, i.e. observed but not present or not observed when present; light red = sporadically observed (regarded as not observed) but present.

Samples	Tissue Type	Observed peaks (x)	Possible peaks (n)	Scoring result	Interpretation
Blood : Saliva (1:1)	Blood	8	3*4	x ≥ n/2	Observed
	Saliva	7	2*4	x ≥ n/2	Observed
	Semen	0	2*4	x = 0	Not observed
	Menstrual secretion	0	3*4	x = 0	Not observed
	Vaginal mucosa	3	3*4	x ≤ n/2	Sporadically observed, no reliable statement possible
	Skin	1	3*4	x ≤ n/2	Sporadically observed, no reliable statement possible
	Mucosa	4	1*4	x ≥ n/2	Observed and fits with saliva
Blood : Saliva : Skin (1:1:1)	Blood	10	3*4	x ≥ n/2	Observed
	Saliva	8	2*4	x ≥ n/2	Observed
	Semen	0	2*4	x = 0	Not observed
	Menstrual secretion	2	3*4	x ≤ n/2	Sporadically observed, no reliable statement possible
	Vaginal mucosa	1	3*4	x ≤ n/2	Sporadically observed, no reliable statement possible
	Skin	12	3*4	x ≥ n/2	Observed
	Mucosa	4	1*4	x ≥ n/2	Observed and fits with saliva
Vaginal : Semen (fertile) (10:1)	Blood	2	3*4	x ≤ n/2	Sporadically observed, no reliable statement possible
	Saliva	0	2*4	x = 0	Not observed
	Semen	7	2*4	x ≥ n/2	Observed
	Menstrual secretion	3	3*4	x ≤ n/2	Sporadically observed, no reliable statement possible
	Vaginal mucosa	8	3*4	x ≥ n/2	Observed
	Skin	6	3*4	x ≥ n/2	Observed
	Mucosa	4	1*4	x ≥ n/2	Observed and fits with vaginal mucosa
Menstrual : Semen (sterile) : Skin (10:5:1)	Blood	6	3*4	x ≥ n/2	Observed and fits with menstrual secretion
	Saliva	0	2*4	x = 0	Not observed
	Semen	4	2*4	x ≥ n/2	Observed
	Menstrual secretion	12	3*4	x ≥ n/2	Observed
	Vaginal mucosa	8	3*4	x ≥ n/2	Observed and fits with menstrual secretion
	Skin	4	3*4	x ≤ n/2	Sporadically observed, no reliable statement possible
	Mucosa	4	1*4	x ≥ n/2	Observed and fits with menstrual secretion

Complex mock casework

To further investigate the applicability of the 20plex and the 'x=n/2' counting system, four challenging mock casework stains (rather than cDNAs) were also analysed. Potential variations between different laboratories could help improve the system for future application of mRNA profiling in 'real' forensic cases. The methodology used was the same as for the previously presented body fluid stains; therefore, results can be easily compared. The serial cDNA input approach was followed and an optimal cDNA amount was determined prior to generating four replicate mRNA profiles per sample. In addition to the differentiation of the tissue of origin, the number and genders of the donors were also estimated through DNA profiling of the co-extracted DNA. The organising laboratory also asked for a final statement determining which cell types

were thought to be present or absent. Taking into account the results so far using the 20plex system, the markers LOR and HBD1 were excluded during scoring because of either non-specific signals with vaginal fluid (LOR) or insufficient amplification (HBD1).

For stain 1 (mother and daughter contributing saliva), both saliva and menstrual secretion were scored as "observed" [Table 4-9]. The false menstrual blood score is believed to be related to the analysis of slightly overloaded mRNA profiles (obtained optimal cDNA input=4 µl) in which trailing signals are found about ten nucleotides before the parent peaks. The trailing signal of saliva marker HTN3 fitted within the bin of the MMP7 marker (0.3 nucleotides smaller), however, there were also some background signal for MMP10. Housekeeping signals also showed these trailing signals. These false positive events can be corrected by re-designing the PCR primers and adjusting the PCR product length; however, it is believed that gaining more expertise in analysing mRNA profiles would result in successfully recognising these events. The peaks for the saliva markers were much higher; therefore, it could be concluded that the saliva was the main contributor to the sample. On the other hand, no false positive scoring occurred for stain 2 (female menstrual blood/skin and male blood). However, due to the co-expression of markers in peripheral and menstrual blood, the identification of blood was masked under the "observed and fits" category. Even though almost half of the total rfu weight in the obtained DNA profiles corresponded to the blood donor, it seemed impossible to assign donors to body fluids since menstrual blood gives relatively low DNA signals while blood results in lower mRNA signals (Harteveld *et al.*, 2013) [Figure 4-7].

Furthermore, stain 3 (skin and diluted blood from one female as well as skin from another female) contained low amounts of cell material (DNA concentration of 0.16 ng/µl) and proved to be challenging. Only five out of nine laboratories scored skin as "observed" and only one managed to detect blood. It was suggested by the organising laboratory that ethanol precipitation of mRNA prior to cDNA synthesis would potentially solve the problem of low mRNA signals. Finally, for stain 4 (nail clipping from a male, vaginal mucosa of a female, azoospermic semen from another male), the seminal fluid was completely undetected since no mRNA peaks were observed. Since there were no spermatozoa, no DNA peaks were observed. Since SEMG1 is the only

marker that is expected to be amplified, it had to be present in all four replicates to reach the 'x≥n/2' level. Taking into account the above results, it was concluded that the "sporadically observed (& fits)" category was very helpful since it lowered the number of false positive results caused by non-specific signals.

Table 4-9. Identification results of complex mock casework samples

Stains	Tissues	Observed	Observed & fits	Sporadically obs. & fits	Not observed	Spor. obs. not reliable	Not specific (overload)
1	Blood		x				
	Saliva	x					
	Semen				x		
	Skin					x	
	Menstrual secretion	x					
	Vaginal mucosa			x			
	General mucosa		x				
2	Blood		x				
	Saliva					x	
	Semen				x		
	Skin	x					
	Menstrual secretion	x					
	Vaginal mucosa			x			
	General mucosa			x			
3	Blood					x	
	Saliva				x		
	Semen				x		
	Skin	x					
	Menstrual secretion					x	
	Vaginal mucosa				x		
	General mucosa				x		
4	Blood				x		
	Saliva				x		
	Semen				x		
	Skin	x					
	Menstrual secretion					x	
	Vaginal mucosa	x					
	General mucosa		x				

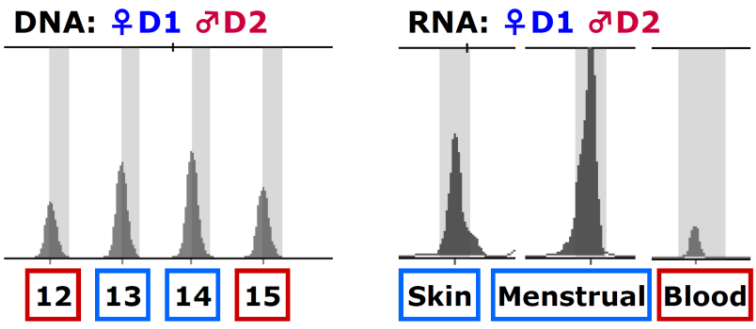


Figure 4-7. DNA and RNA signal comparison in stain 2 consisting of two donors and three cell types (♀D1 skin and menstrual blood, ♂D2 blood)

Data interpretation exercise

The overall verbal conclusions given by all laboratories regarding the complex casework samples were so variable suggesting that guidelines on reporting final mRNA results are also needed. Therefore, the same set of DNA and mRNA profiles (produced by the organising laboratory) of the four complex stains was asked to be interpreted by all laboratories. The interpretation of DNA results, the conclusions regarding the estimated minimum number of contributors as well as the donors' gender were all correct. For the mRNA profiling results, two stains were interpreted correctly while the other two led to false conclusions.

The issues regarding stain 1 (false identification of menstrual blood) was resolved and the overall interpretation *"The sample is more likely to be a mixture of two females; saliva was the only body fluid detected so it belongs to either the major donor only or to both of them"* was correct. On the other hand, for stain 2, blood was still masked during scoring and no statement could be made regarding its presence in the sample. Since strong HBB and CD93 peaks were present in all replicates, it should have been reported as observed. The statement *"The sample is more likely to be a mixture of a female contributing menstrual secretion and a male contributing skin"* was incorrect and misleading as the presence of blood was missed. This wrong classification was also shared by other laboratories, highlighting the high possibility of missing blood when menstrual blood is present.

However, for stain 3, blood was reported as "observed" since eleven out of twelve blood-specific peaks were present in the mRNA profiles. Even though no statement could be made regarding the presence of menstrual blood (three peaks present), the final conclusion *"The sample is more likely to be a mixture of two female donors contributing blood and skin; however, no association can be made between donors and body fluids"* was correct. Finally, all three contributing tissues were identified in stain 4 and the assumption of the presence of azoospermic semen could be made. However, the wording in the final statement *"The sample is more likely to be a mixture of a female contributing vaginal fluid and an azoospermic male as his semen was detected but no DNA profile was obtained. Skin cells were also present but no statement can be made regarding its donor"* was incorrect since there was biological material from two males.

4.4 Final remarks

As discussed above, mRNA profiling could be a very useful tool for the identification of forensically relevant body fluids and tissues. A number of multiplex PCR assays have been developed such as the tissue-specific EDNAP multiplexes or the revised TissueID 20plex (developed by the NFI) that allows for simultaneous identification of biological fluids. However, extensive validation was required to assess the overall performance of these assays prior to applying such a test in forensic casework. In general, all analysed assays were found to be extremely sensitive and able to accurately determine the tissue source of a stain from minute amounts of starting material; nevertheless, for each body fluid certain mRNA markers were found to be more robust. A set of freshly prepared, aged, degraded or mixed body fluid stains together with selected complex mock casework were used to further assess the applicability of mRNA profiling and to identify any flaws in the overall assay performance.

False negative results were usually obtained in cases where samples were of very low quantity or because of suspected inter-individual variation in gene expression levels. On the other hand, false positives were occasionally seen as a result of the low expression of cell type-specific mRNAs in non-target tissues. Overall, it was observed that the identification of vaginal material was the most challenging task since its gene expression pattern often resembled those of saliva and skin, both tissues of epithelial origin. In most cases skin mRNA markers were detected in vaginal secretion; however, it cannot be excluded that this could also be due to contamination with skin while sampling. As a whole, the expression of menstrual blood- and vaginal fluid-specific markers can be affected by the menstrual cycle and individual differences in menstrual flow but it has not yet been experimentally tested. Furthermore, the use of vaginal-specific bacterial mRNA markers should be followed with caution since co-expression of certain bacteria (such as Ljen) was obtained in other tissues.

Moreover, technical aspects such as the cDNA input in PCR was found to affect the resulting mRNA profiles, since overloaded mRNA profiles contained trailing signals that fell within other markers' bins. These findings strongly support the need for adjusting the cDNA amount prior to generating a set of PCR replicates. They also suggest the potential advantages of applying a scoring system as well as the set-up of specific guidelines in regards to the interpretation of mRNA profiles. The analysis of

mixtures revealed that it is possible that some markers are occasionally detected only in some of the replicates; therefore, by using a scoring system and including the category "sporadically observed (& fits)", false positives could be avoided. This variability of markers presence amongst replicates is significantly enhanced when the expression (translated as peak heights) is low and is believed to be due to experimental variation (mainly PCR). Additionally, it is essential that not only the presence or absence of a peak, but also the strength of each mRNA signal is considered so that sporadic detection of markers can be better assessed. Especially in the case of a menstrual blood/blood mixture, blood could be easily disguised since the detection of blood-specific markers is common in menstrual blood. It can be generally accepted that it is almost impossible to compare DNA and mRNA profiles in an attempt to correlate donors and tissues, since DNA and RNA peaks do not demonstrate the same signal strengths.

The observed variations in mRNA markers detection can be used to improve the existing 20plex assay. Adjusting the primer concentrations could lead to better and more even amplification across the markers. For example, the vaginal marker HBD1 was insufficiently amplified within the multiplex system so either the experimental conditions could be adjusted or the marker could be replaced by the highly specific marker MYOZ1. Similar suggestions could be made for the blood marker AMICA1 and the housekeeping marker GAPDH. The skin marker LOR often resulted in false positive identification of skin so it should be removed or replaced. Furthermore, the mucosa marker KRT4 was not found to be particularly useful since there were always other cell-type specific markers amplified when positive KRT4 mRNA signals were obtained. In order to avoid false positive identification of menstrual blood in saliva stains, the size of the MMP7 fragment should be adjusted to avoid overlap; re-design of the primers would be recommended. It was noticed that in cases where azoospermic semen was present in the sample, the existence of the seminal fluid-specific SEMG1 marker was required in all replicates in order to reach the 'x=n/2' threshold for "observed". Consequently, the addition of an extra seminal marker would be beneficial to prevent false negative results. Finally, alongside mRNA profiling, DNA profiling was also carried out. The employed DNA/RNA co-extraction method was found to be efficient and independent of the starting biological material. Most body fluids stains resulted in full STR profiles that could be used in identifying the donor.

5 Identification of body fluid-specific differentially methylated CpG sites

5.1 Introduction

As shown in Chapter 3, tissue-specific differentially methylated CpG sites have already been investigated for use in forensic tissue identification (Frumkin *et al.*, 2011; Madi *et al.*, 2012). Although most of these techniques seem to be quite sensitive and therefore, suitable for use in forensic specimens, marker specificity issues have been reported. With the exception of semen, the discovery of highly specific methylation markers for most analysed tissues including saliva or vaginal fluid has proven a challenging task. Consequently, it is important that more potentially tissue-specific markers are investigated.

5.1.1 Approaches for identifying suitable CpG sites

In general, there are several approaches that can be used in discovering tissue-specific differentially methylated CpG sites, each demonstrating their own advantages and drawbacks. Firstly, large-scale discovery experiments analysing DNA from various tissues of the same or a group of individuals could reveal tissue-specific methylation patterns. Genome-wide or chromosome-specific methylation studies have already been published in literature such as the study by Rakyan *et al* (2008) analysing a total of 13 different tissues. Therefore, similar genome-wide experiments can be performed using forensically relevant tissues. This approach allows for the investigation of thousands of CpG sites at once, thus maximising the chance of detecting potential methylation differences. Also, CpG sites located in various parts of the genome are analysed, therefore this approach does not limit researchers in CpG sites located in gene promoter regions only. Nevertheless, such methods are costly and data analysis usually requires expertise in bioinformatics. Evaluating existing analysed methylation data could potentially overcome some of these challenges; however, it is difficult to obtain such data for all forensically relevant tissues (e.g. menstrual blood).

Furthermore, since DNA methylation is known to regulate various genome functions including gene regulation, a second approach includes the investigation of already reported genes showing tissue-specific gene expression or DNA methylation patterns. As an example, even at the early stages of epigenetic research globin genes were reported to be regulated through differential DNA methylation patterns (Bartzeliotou & Dimitriadis, 1989; Shen & Maniatis, 1980). Also, as shown in Chapter 4, forensic

researchers have identified several mRNA molecules showing tissue-specific gene expression. Therefore, the methylation status of the corresponding genes could be examined to assess if the reported tissue-specific gene expression is due to differential methylation patterns. This approach seems very promising and could lead to the identification of highly specific markers. However, since current methodologies such as methylation specific PCR or bisulphite Pyrosequencing® permit the analysis of only a subset of CpGs, this process could be time-consuming. Also, by employing a promoter-focused approach, potential methylation markers outside gene regions would be overlooked.

Lastly, as mentioned earlier in section 3.3, researchers have also identified cell type-specific differentially methylated markers rather than tissue-specific ones (Baron *et al.*, 2006). In their study, the authors used CDMs for precise and robust quantification of subpopulations in cell cultures, in particular immune cell types in whole blood. However, it is believed that CDMs could also be used to differentiate between tissues since each tissue contains a different ratio of immune cell types. Although this approach offers advantages, proposed methods require large amounts of starting material (for example, at least 100 µl of whole blood). As an example, the principle of published qPCR assays for immune cell-specific methylation detection is presented in Figure 5-1. One should also take into account that any medical condition leading to immunological responses such as infectious diseases, allergic responses, autoimmunity and cancer could alter the results.

5.1.2 Selection of the appropriate methodology

As mentioned in section 1.1.5, depending upon the application and number of CpG sites of interest, different methodologies for methylation detection can be applied including treatment of DNA with sodium bisulphite, use of methylation-sensitive restriction enzymes and use of proteins that interact with methylated cytosines. The advantages and drawbacks of each approach were discussed and it was concluded that bisulphite treatment was perhaps the most appropriate in the context of forensic analysis. This is due to the fact that DNA integrity is maintained during analysis (in contrast with the use of restriction enzymes) and the quantity of starting DNA material needed seems to be relatively low (compared to the methylated DNA immunoprecipitation).

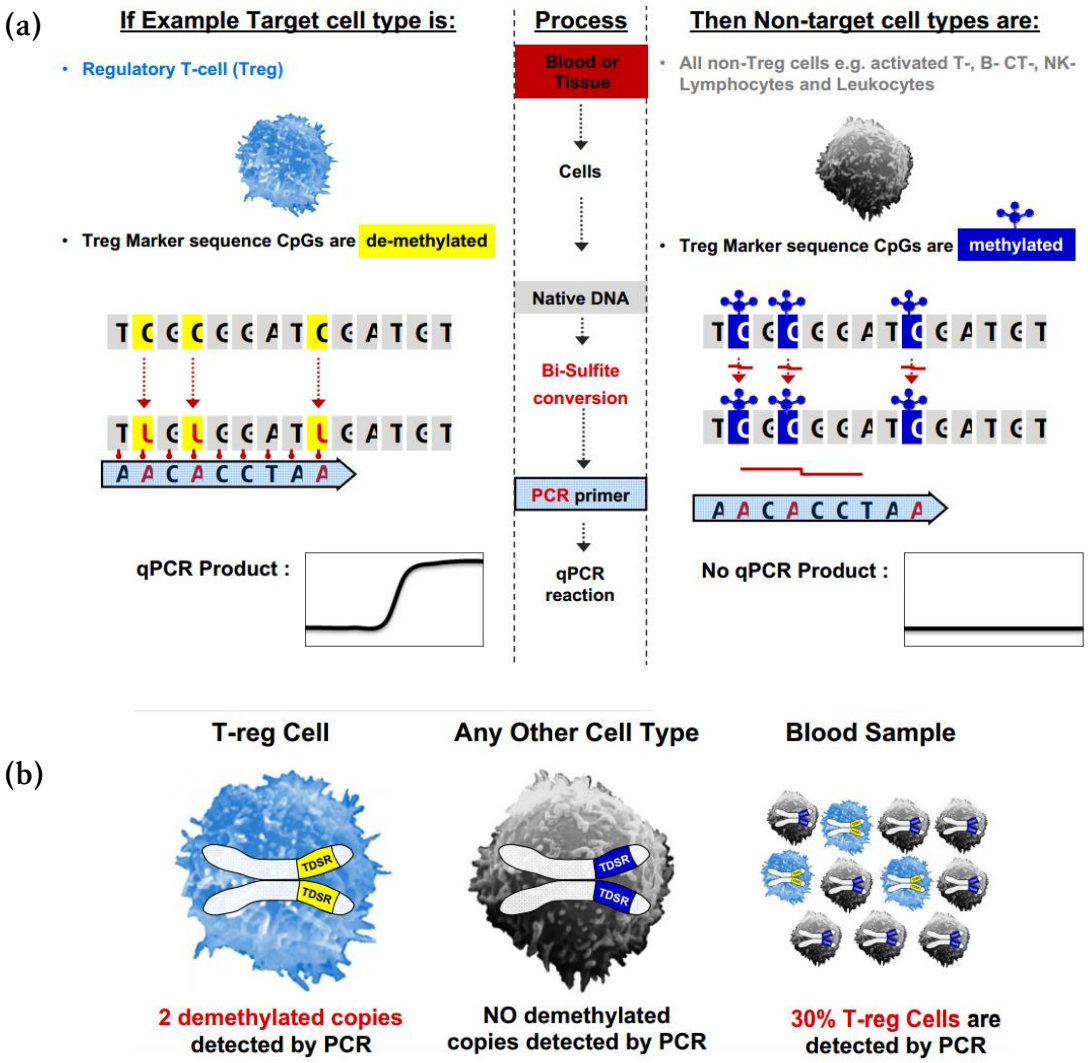


Figure 5-1. Immune cell-specific methylation detection via qPCR assays (confidential data from Epiontis, 2013)

(a) Overview of developed methodology to detect Treg-specific methylation. (b) Methylation values correspond to number of cells; therefore they are useful for cell counting.

So far, there have been various reported methods based on bisulphite treatment of DNA, which will be presented here. The most common one, methylation-specific PCR (MSP), is a standard amplification reaction performed using two different sets of primers, which are either specific for methylated or unmethylated DNA respectively (Herman *et al.*, 1996). As a result, the binding site must include the CpG site(s) of interest. When analysing multiple CpG sites, validation of each CpG site needs to be performed individually. Preferential primer binding may also introduce a degree of bias that needs to be taken into consideration. MSP mainly provides qualitative data leading to either a positive or negative result and has given contradictory results in some studies (Martin-Subero *et al.*, 2009). This is especially true for samples demonstrating an intermediate methylation profile. Likewise, methyLight is a standard protocol based on TaqMan technology with the difference that the primers

are specific for bisulphite converted DNA (Eads *et al.*, 2000). This approach allows for the separate analysis of differentially methylated sequences that comprise the binding site of the probe used.

In combined bisulphite restriction analysis (COBRA), bisulphite treated DNA is used for amplification by a primer set flanking the recognition site of a restriction endonuclease (Xiong & Laird, 1997). An advantage of this method is the use of restriction enzymes that are not methylation-specific, however, the analysis is restricted to single CpGs. Similarly, methylation-sensitive single nucleotide primer extension (MS-SNuPE) is another technique that can be applied (Gonzalgo & Jones, 2002); the main drawbacks associated with this method include the restriction of the primer location as well as the use of radioactive nucleotides posing a health risk. On the other hand, high-resolution melting analysis (HRM) enables rapid quantification of methylation changes in genomic loci but does not provide information about the exact location of the methylated/unmethylated CpG sites (Wong & Dobrovic, 2011). Differentially methylated amplicons do not behave in a similar way and could affect analysis. Lastly, in bisulphite sequencing (BS), PCR products that have been amplified using bisulphite converted DNA are either sequenced directly or first cloned into a suitable vector and subsequently transferred into a bacterial genome (Clark *et al.*, 1994). However, BS is not generally considered suitable for quantitative analysis.

Most standard methylation techniques give only qualitative or semi-quantitative results, which can lead to inaccurate interpretation; however, absolute quantification allows for the differentiation of physiologically meaningful methylation changes. It is believed that the sensitivity and accuracy of Pyrosequencing® surpass alternative methods [Figure 5-2], thereby making its application ideal in forensic science, e.g. analysis of autosomal STR markers (Divne *et al.*, 2010), s equencing of mitochondrial DNA (Andreasson *et al.*, 2007) and Y-chromosomal STR analysis (Edlund & Allen, 2009). Pyrosequencing® technology refers to the real-time pyrophosphate detection of DNA sequencing. It is a simple, robust and sensitive technique, which enables accurate and quantitative analysis of DNA sequence variation and has demonstrated unrivaled performance in determining the methylation status of individual or multiple adjacent CpG sites (Dupont *et al.*, 2004). Overall, Pyrosequencing® allows for single nucleotide polymorphisms (SNPs), mutation and CpG analysis.

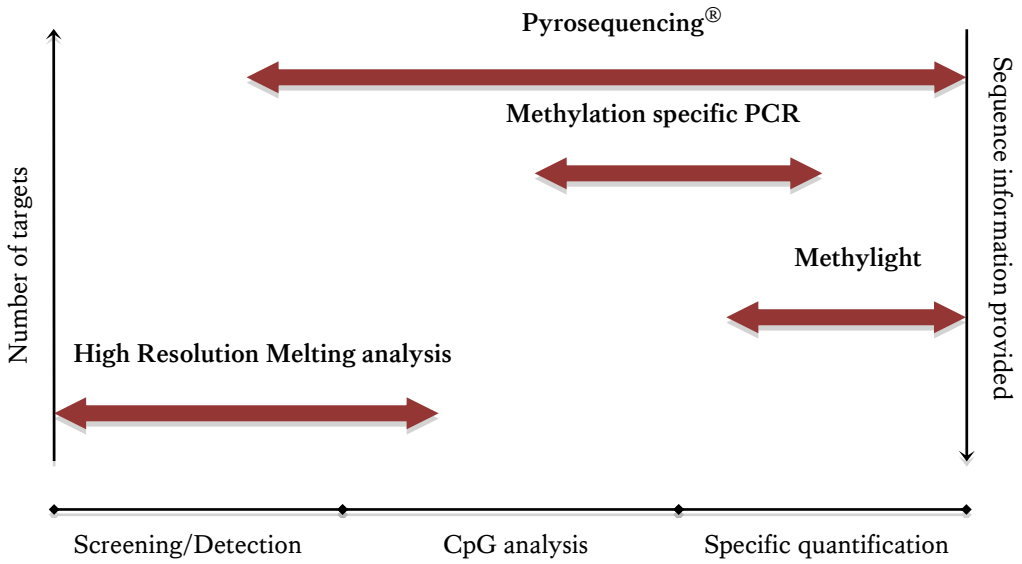


Figure 5-2. Schematic comparison of DNA methylation techniques

Pyrosequencing® can reliably determine the nucleotide sequence starting with the first base following the primer. Although the read length is limited in comparison to other techniques (100-150 bp), the use of a single-stranded binding protein has been reported to efficiently allow for longer sequences to be analysed (Ronaghi, 2000). Assay design is very versatile as there is no restriction on primer locations; however, suitable, non-CpG-containing primers are essential to avoid possible PCR bias. Nevertheless, the use of standards with known methylation levels can be used to normalise the quantification procedure. Furthermore, extra precautions are also needed since non-specific signals are sometimes generated due to the low reaction temperature of the sequencing process (28 °C). This problem can be easily overcome with the use of quality controls such as dead injections, reference peaks and built-in bisulphite conversion controls.

In a forensic setting, quantitative and reliable estimation of DNA methylation levels pose a major challenge; however, recent advances allow for the sensitive detection of DNA methylation at multiple CpG sites from less than 1 ng of DNA from trace amounts of body fluids (Paliwal *et al.*, 2010; Xu *et al.*, 2012). Paliwal *et al* proposed a method that involves genome-wide amplification of bisulphite-modified DNA template and quantitative methylation detection using Pyrosequencing® starting with extremely low DNA amounts (100 pg). Genome-wide amplification can overcome the limitation of the minute quantities of DNA, however, good quality and quantity of starting DNA material remains the most critical determinant of such an experiment.

Furthermore, the possibility of any methylated/unmethylated allele bias cannot be formally ruled out. In another study, Xu *et al* recommended a modified, more efficient method of bisulphite genomic DNA sequencing from dried blood spot microvolume samples. After bisulphite treatment, the methylation status of the differentially methylated region (DMR) of the maternally imprinted gene named small nuclear ribonucleoprotein polypeptide N (*SNRPN)* was investigated using methylation-specific PCR and direct sequencing.

5.1.3 Aim and Objectives

The aim of this study was to investigate selected body fluid-specific differentially methylated CpG sites and assess the possibility of using a small number of DNA methylation markers to accurately differentiate between forensically relevant body fluids.

In order to meet the aim, the following objectives were addressed:

- A suitable Pyrosequencing®-based method was evaluated using a pre-designed HBA1 assay (QIAGEN). Validation of the method was achieved by establishing sensitivity, accuracy, reproducibility and linearity of methylation quantification.
- Potential tissue-specific differentially methylated CpG sites were identified using three different approaches: (a) by validating reported body fluid-specific markers in the literature, (b) by analysing published methylation data obtained from blood, semen and buccal cells, and lastly (c) via validating proposed immune cell-type specific markers as part of a collaborative project within the EuroForGen consortium.
- Suitable bisulphite Pyrosequencing® assays were then developed and optimised enabling the measurement of the methylation levels of selected CpG sites
- The specificity of the proposed markers was evaluated by analysing a large set of body fluid samples including blood, semen, saliva, menstrual blood, vaginal secretion, skin and urine.
- Promising markers were then further validated by testing their sensitivity and their applicability using mock casework samples.

5.2 Experimental

5.2.1 Samples

5.2.1.1 Fresh body fluid/tissue samples

All body fluid samples were collected from a total of 168 volunteers of both sexes and various ethnic backgrounds with ages ranging from 16 to 70 years old. Individuals had the choice to donate one or more body fluids/tissues including whole blood, saliva, buccal cells, seminal fluid, vaginal fluid, menstrual secretion, skin, urine and nasal fluid. Samples were collected as previously described in section 2.1 and were stored at -20 °C until extraction. Additionally, six nasal blood samples were provided by a collaborative laboratory as part of the EuroForGen project. As shown in Table 5-1, the final dataset comprised of 254 body fluid samples. DNA samples were firstly amplified using the PowerPlex® ESI 16 kit (Promega) as described in section 2.2.4.5 to confirm that they all were of a single source and no contamination had taken place.

Table 5-1. Body fluid samples analysed in this study

Set	Samples	♀/♂	Age range (mean) (years)
Whole blood	65	28/37	16-70 (38)
Saliva	46	31/15	19-34 (26)
Buccal cells	15	12/3	18-50 (32)
Semen	18	0/18	19-34 (25)
Vaginal fluid	22	22/0	20-33 (25)
Menstrual blood	20	20/0	22-33 (26)
Skin	28	20/5	21-33 (25)
Urine	20	10/10	21-60 (27)
Nasal fluid	13	8/5	22-37 (28)
Nasal blood	6	N/A	N/A

5.2.1.2 Aged samples

For validation purposes, a set of aged stains or DNA samples were used which had been previously collected for research purposes. The set included:

- Five blood samples stored for 9-18 years in the dark and at room temperature
- Four semen samples stored for 16 years at -20 °C
- Five blood DNA samples stored for 1-9 years at -20 °C

5.2.1.3 Body fluid stains

Using freshly collected samples, the following body fluid stains were prepared:

- Mock casework blood samples including 1 µl on a piece of jeans fabric, 3 µl on shirt which was washed afterwards, 3 µl on a towel and 5 µl on tissue paper
- Mixed stains including 1 µl blood:1 µl semen and 2 µl blood:1 µl saliva on swabs
- Artificially-degraded blood samples by exposing 1 µl under UV for 0, 10, 30, 60, 90, 120 and 240 minutes
- 5 µl of blood on fabric stored at various temperatures (-20 ^{o}C, 4 ^{o}C, 25 ^{o}C, 37 ^{o}C, outdoor) for a week

All stains were left at room temperature to dry overnight before analysis.

5.2.2 Selection of suitable tissue-associated CpG sites

As previously mentioned, there are various ways in identifying potential tissue-specific differentially methylated CpG sites; here, three different approaches were followed.

5.2.2.1 Reported tissue-associated CpG sites in the literature

Neumann *et al* analysed the methylation status of the embryonal Fyn-associated substrate (*EFS*) gene in various tissues including blood, buccal cells, sperm, and brain and reported tissue-specific methylation patterns (Neumann *et al.*, 2011). In more detail, they found that *EFS* is highly methylated in blood, completely unmethylated in sperm and partially methylated in buccal cells. Therefore, it was thought that this marker could be a good candidate for blood detection; however, other forensically relevant tissues needed to be investigated before any conclusions could be made. Authors analysed a total of eleven CpG sites located in the region Chr4: 23,835,859-23,835,970, which were used for the development of Pyrosequencing® assays.

5.2.2.2 Tissue-specific CpGs via genome-wide methylation data analysis

There are various published studies investigating genome-wide methylation patterns across various tissues that have either revealed tissue-specific (Rakyan *et al.*, 2008; Thompson *et al.*, 2013) or age-associated CpG sites (Rakyan *et al.*, 2010; Zykovich *et al.*, 2014). Combining available data from two studies analysed by different large-scale methodologies for three forensically relevant tissues (whole blood, buccal cells and

sperm), the methylation status of a total of 3,305 CpGs was obtained. Methylation data for seven blood and two sperm samples were obtained from the study by Rakyan *et al* (2008), while methylation values for ten buccal swabs were gathered from Rakyan *et al* (2010).

In order to identify potential CpG sites of interest, the average methylation level of each tissue for every CpG site was obtained. Initially, the criterion of choosing suitable CpG sites was set as a minimum of 70% methylation difference between the tissue in question and the remaining two. For example, a 'good' blood specific marker would be one that showed to be methylated in blood (>80%) and unmethylated in sperm and buccal cells (<10%) or vice versa. As a result, 14 blood-specific, 20 saliva-specific and 365 semen-specific CpG sites were identified. The number of semen-specific CpG sites was relatively high, which could reflect on the different and unique functions of sperm DNA. Thus, only for semen the minimum difference of methylation levels was increased to 85%, thus, the number of potential CpG sites decreased down to 22.

Both tissue-specific methylation and non-methylation were observed, although the latter was observed comparatively more often. For the purpose of this study, four CpG sites demonstrating the highest methylation difference between the selected tissues were selected for each body fluid (two tissue-specific unmethylated and two tissue-specific methylated markers). Therefore, a total of twelve potentially tissue-specific CpG sites were chosen for Pyrosequencing® assay design [Table 5-2]. Most of them belong to protein-coding genes and are usually located within their 5' end.

5.2.2.3 Immune cell-specific CpG sites

As part of a EuroForGen collaborative project, it was thought that body fluids can be successfully differentiated through specific immune cell infiltrates. Initial studies using qPCR-based assays revealed a set of potential immune cell-specific methylation markers [Figure 5-3]. For this study, six loci specific to the following cell types involved in the immune response: T cells, neutrophils, NK cells and naive CD8+, CD8A+ and CD8B+ T cells were obtained for Pyrosequencing® assay design [Table 5-3].

Table 5-2. Identified blood-, semen- and buccal cell-specific methylation markers resulting from analysing genome-wide methylation data (Rakyan *et al.*, 2008 & Rakyan *et al.*, 2010)

Essential information regarding chromosomal location of the selected CpG sites, the gene they belong to as well as their exact position within the gene. The mean methylation levels obtained by the genome-wide analysis was used to calculate the minimum methylation difference.

Marker	Chromosomal location	Gene	Position within gene	Mean DNA methylation levels			Difference
				Buccal cells	Blood	Sperm	
cg05761971	2:38,177,677	FAM82A1	5' Upstream sequence	10,8%	87,2%	88,2%	76,4%
cg16779976	10:98,031,125	BLNK	Exon 1	10,5%	85,6%	88,4%	75,1%
cg15731815	1:6,269,260	RNF207	Intron 5	84,3%	7,9%	7,7%	76,4%
cg08258650	11:35,441,900	RP4-683L5.1	Intron 1	77,1%	5,5%	5,0%	71,5%
cg17518965	19:3,178,955	S1PR4	Exon 1	87,7%	2,0%	95,5%	85,7%
cg26285698	16:29,757,334	C16orf54	5' Upstream	86,1%	8,5%	87,2%	77,6%
cg13763232	3:14,443,428	SLC6A6	5' Upstream	14,5%	90,3%	6,0%	75,8%
cg21613754	2:219,133,847	AAMP	Intron 2	14,5%	86,1%	12,6%	71,6%
cg04382920	4:128,651,526	SLC25A31-001	5' Upstream	99,0%	97,0%	2,0%	95,0%
cg11768416	12:91,348,697	C12orf12	Exon 1	98,1%	95,8%	1,9%	93,9%
cg01318557	7:73,624,319	LAT2	Exon 1	1,9%	4,3%	95,6%	91,3%
cg05656364	2:85,804,732	VAMP8	Exon 1	1,5%	5,3%	90,3%	85,0%

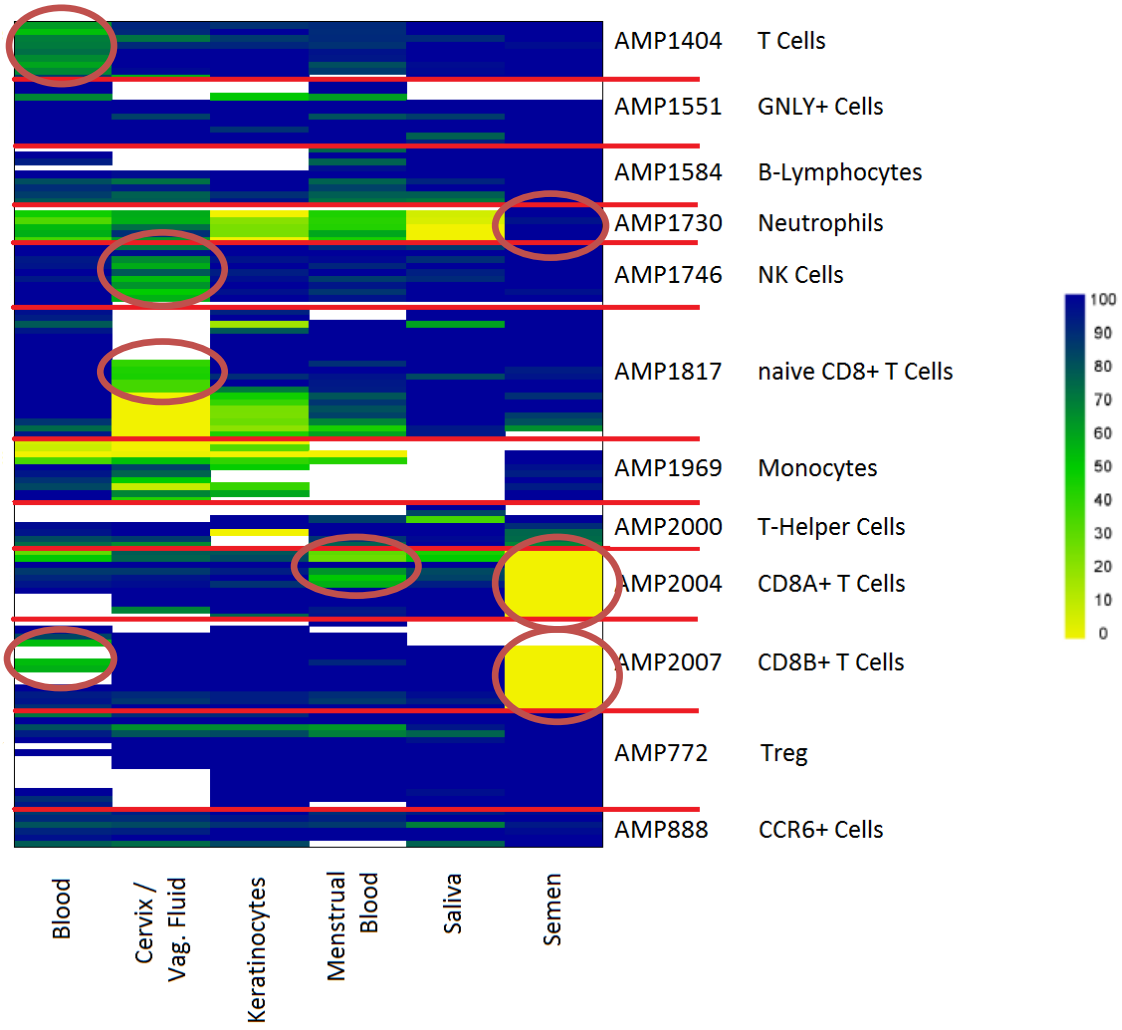


Figure 5-3. Epigenetic profiles of various immune cell markers in forensically relevant body fluids (confidential data from Epiontis 2013)

A total of six samples from six different tissues were analysed with various qPCR assays. Each column corresponds to a different tissue, while each row separated by the horizontal red lines represents an assay corresponding to a specific cell type (e.g. AMP1404 measures methylation levels of T cells). The horizontal lines within each row correspond to different CpG sites; each assay can measure numerous CpG sites at once. Methylation levels are presented using a colour scale (0%-yellow, 100%-blue). Red circles indicate potential immune cell-specific differentially methylated CpG sites.

Table 5-3. Selected loci for analysing immune cell-specific methylation levels

Loci	Immune cell type	Chromosomal position	Gene
AMP1404	T cells	11:118,213,614-118,214,043	CD3
AMP1730	Neutrophils	9:130,911,530-130,911,888	LCN2
AMP1746	NK cells	11:3,121,298-3,121,543	OSBPL5
AMP1817	naive CD8+ T cells	11:66,083,367-66,083,805	CD248
AMP2004	CD8A+ T cells	12:6,899,890-6,900,346	CD4
AMP2007	CD8B+ T cells	2:87,048,390-87,048,818	CD8b

5.2.3 Pre-designed PyroMark CpG assay (HBA1)

For the purpose of the initial Pyrosequencing® assay development, a PyroMark CpG assay capable of analysing four CpG sites belonging to the haemoglobin, alpha 1 (*HBA1*) gene was available and chosen for this study. Following the PCR conditions mentioned in section 2.2.4.3, the amplicon has a length of 150 bp and the analysed sequence is the following: CCCGCGCTGCAGGCGTCG (Chr16: 231,036-231,054).

5.2.4 Bisulphite Pyrosequencing® assay design

Following the guidelines regarding assay design mentioned in sections 2.2.4.1 and 2.2.6.1.1, bisulphite Pyrosequencing® protocols were designed using the BiSearch primer design tool. Each assay includes a 10X PCR primer set (forward and reverse) as well as 10X sequencing primer(s). For some assays, two sequencing primers were used in order to accommodate all CpG sites of interest. For EFS, ten out of the reported eleven CpG sites were investigated due to location and assay design restrictions. Furthermore, a total of twelve assays were designed to investigate the CpG sites identified via the analysis of genome-wide methylation data. However, assay design allowed for the co-analysis of a few adjacent CpG sites per assay resulting in a total of 48 analysed CpGs. Therefore, to simplify data analysis, these assays were named according to the tissue they were specific for (BL- for blood, SE- for semen and BU- for buccal cells) and if they were expected to be unmethylated (-U-) or methylated (-M-) in that specific tissue. As an example, the assay BLM1 (cg13763232) indicated that the investigated CpG site is expected to be methylated in blood, while being unmethylated in semen and buccal cell samples. Lastly, regarding the immune cell markers, only the CpG sites targeted by primer/probes in the previously developed Taqman-based assays were investigated (39 CpG sites in total resulting in six Pyrosequencing® assays). Information regarding all 19 designed assays measuring the methylation status of a total of 97 CpG sites is presented in Tables 5-4 and 5-5.

5.2.5 Optimised bisulphite PCR conditions

In initial experiments, the AmpliTaq Gold PCR master mix (QIAGEN), often used in forensic DNA profiling as well as the PyroMark PCR (QIAGEN), which is specifically designed for Pyrosequencing® analysis were utilised. However, the PCR

efficiency was very low and non-specific products or primer-dimers were observed in most assays (data not shown). Consequently, it was made clear that a more 'specific' DNA polymerase was needed to overcome potential difficulties when amplifying 'difficult' and of reduced complexity DNA such as bisulphite-converted DNA sequences.

As mentioned in section 2.2.4.4, ZymoTaq premix (Zymo Research) contains a heat-activated, "hot start" DNA polymerase, which reduces the occurrence of non-specific product or primer-dimer formation when amplifying bisulphite-converted DNA. Starting with the standardised PCR conditions previously described in sections 2.2.4.2 and 2.2.4.4, bisulphite PCR assays were optimised. Briefly, each PCR reaction consisted of 12.5 µl of ZymoTaq PreMix, 1 µl of 25 mM $MgCl_2$ for a final concentration of 2.75 mM (since the ZymoTaq™ Premix also contains 1.75 mM $MgCl_2$), 1µl of each PCR primer (for a final concentration of 0.4 µM), 1 µl of bisulphite DNA template and 8.5 µl of nuclease-free water, for a total reaction volume of 25 µl. The thermocycling program used was: 95 °C for 10 minutes, followed by 45 cycles of 94 °C for 30 seconds, T_m for 30 seconds (SEU1 - 48 °C, EFS, BLU1, AMP1817 - 50 °C, AMP2007 - 52 °C, AMP1404, AMP1730, AMP1746, AMP2000 - 54 °C, BLM1, BLU2, SEM1, BUM1, BUM2, BUU1, BUU2 - 55 °C, SEM2 - 57 °C and SEU2 - 61 °C), 72 °C for 30 seconds, and a final extension step of 72 °C for 7 minutes. Specifically for BLU1, BUM1, BUM2, BUU1 and BUU2, the annealing and extension step of each cycles were 40 seconds long. Following amplification, the quality of PCR products was assessed on a 2% agarose gel as described in section 2.2.5.1.

Table 5-4. Designed bisulphite PCR assays

Essential information regarding the designed bisulphite PCR assays are shown, such as the number of analysed CpGs, the score assigned by the primer design software (the lower the score the more efficient amplification), the primer sequences and length (F for forward and R for reverse), the % G and C content, the melting temperature (T_m), the number of converted cytosines included in the primer sequence (highlighted in red) as well as the length of the final PCR product.

Assay	CpG sites	Score	Primer Sequence (5' → 3')	Length (bp)	%GC	T_m (C°)	Converted Cs	PCR Product (bp)
EFS	10	25.4	F GGTTTTTTTTTTATTAGTTT	20	15.0	55.2	9	195
			R CTTCATATTATCACTAAAACC	21	28.6	54.5	5	
BLM1	4	11.11	F TAGTTGATATTGGTTTGGTA	20	30	55.3	5	159
			R CAAATAACTCAATTTCTCTAC	21	28.6	54.7	3	
BLM2	2	26.27	F AAGTGTTGGGATTTTAGGAGT	21	38.1	60.1	2	180
			R CCTCTTAATTTTCTTTTAAAAAC	23	21.7	58.1	1	
BLU1	3	43.64	F GGTTTATTGTTTTGTATTAT	20	20	53.4	6	127
			R AAATTCTCCAACACCACC	18	44.4	57.4	2	
BLU2	5	13.17	F GAGTTATTTTTTTTGGTGTTGGAT	24	29.2	60.7	8	188
			R ACATCCCCTTAAATTACTTT	20	30	57.2	4	
SEM1	3	20.57	F ATGATTTAGTGGTTGGTAGGAA	22	36.4	59.6	3	147
			R AACACCCCTAAAAAAAAC	18	33.3	57.1	7	
SEM2	3	18.77	F AGTAAGTAGGAAGTGAATTGA	21	33.3	57.1	3	89
			R ATATCTCAAAACAACCCAAA	20	30	56.6	6	
SEU1	10	41.17	F TTTTTATTAGAAAGTTTAGG	19	21.1	53	7	280
			R ACAACAATAACTAAAAATAAATAC	24	16.7	56.4	6	
SEU2	5	36.54	F GGAGGTTGTTTTTTTTTGGTTT	23	30.4	62.4	6	134
			R CTACCAACACCTTCCTCC	18	55.6	58.9	1	

BUM1	6	45.49	F GTAGAGTTTTATTTTTTGTT	20	20	54.4	7	357
			R CTCCTCCACCATAACCTA	18	50	56.9	3	
BUM2	4	28.1	F TAGAGATAGATGGGTTTG	19	36.8	54.8	3	112
			R CTAAATTCCTACAATATTCC	20	30	54.2	4	
BUU1	1	18.15	F GAAAGGTGAGTTATAGAATAGTT	23	30.4	57.9	3	198
			R CAAAATAAATCTCTCCCTT	19	31.6	55	2	
BUU2	2	19.99	F TTGAGATGTTATAAGAGTATTGG	23	30.4	56.8	5	196
			R ACTACTCCCTAAAAAAAC	18	33.3	55.2	7	
AMP1404	5	14.65	F ATGGAAAAGGTGGTTTGAATTT	22	31.8	60.4	3	163
			R TCCTAAAAAAAATCAACCTC	20	30.0	57.0	4	
AMP1730	3	16.97	F AGAAAGAAATAGTATAAGGAAGG	23	30.4	**58.4**	3	182
			R ACCAAAACCAAAAATATAAAACA	23	21.7	**59.0**	9	
AMP1746	7	16.42	F TGAGTTTTTTTATTTTTGTAGTGG	23	26.1	**58.7**	7	137
			R AATCCTACTATCTCTTATCTCTA	23	30.4	**56.2**	6	
AMP1817	11	36.1	F GGTAATTTGGTATTTATTT	19	21.1	53.1	6	188
			R GCTACACTAAAAACTTCC	18	33.3	54.2	5	
AMP2004	5	18.45	F TTTTGTTTTTGTGGTGGATA	20	30.0	**57.4**	6	221
			R ACTCCCACATTAAAAAAAAAA	22	22.7	**59.1**	6	
AMP2007	7	22.38	F GTTAAAGGAGTTGTTAATATTT	22	22.7	**56.5**	4	165
			R GCAAAACAAAACCCCATATT	20	30.0	**58.0**	5	

Table 5-5. Pyrosequencing® DNA methylation assays

The table includes details of the sequenced chromosomal locations, the sequencing primer sequences as well as the nucleotide dispensation order. The CpG sites in question are highlighted in grey, bisulphite-conversion controls in bold while dead injections are underlined.

Assays	Chromosomal location	Sequencing primer (5' → 3')	Sequence to be analysed (5' → 3')	Dispensation order
EFS	14: 23,835,857-91 14: 23,835,902-51	TTGTTTTTTTTTATGGGAGGG GGTTTTTAGTAGAGTTTTTTTA	AGCGCGCGCCTTCCGCCCAGCGGGGCCCTTAGCAG ATCCTCGGCGCCTCCCCTACACAGGGTTCGCTGGGCCGTTCTTGCGGGGC	GATGTCGTCGTCGTCGTCATGTCGTCAGTCAG GATCGTCGTCATCATCAGTCGTCAGTCGTCAGTCGTC
BLM1	3: 14,443,420-47	ATATTGGTTTGGTATTTTTTAGAT	GCCGGGTGCGTTCCGAGTTCTCTACGCT	AGTCGTGATCGTCGAGTCAGTCGT
BLM2	2: 219,133,845-61	GTTGGGATTTTAGGAGTGAGTTAT	CACGCCCGGCCAAAGAT	ATCAGTCGATCGATCAGAT
BLU1	19: 3,178,931-59	GGTTGGAGGATGGCGGTTTGGG	GGCCCTGCGGGGGCTGTCGGTGGCCGCCA	AGTCGTCGATCGTCGTGTCGTCA
BLU2	16: 29,757,190-227	AGGTGTAGGTTTGGAGGGTGTAGG	GCGGTCCGCCTCTCAGACGTAGAGGCCCGGCCTCGGAT	AGTCGATCGTCAGATCGTAGAGTCGATCGAT
SEM1	7: 73,624,318-35	GATTTAGTGGTTGGTAGGAAGT	CCGCCCTGCCCGCCCGCC	ATCGTCGATCGATCGATC
SEM2	2: 85,804,723-67	ATTGAGGGTTATTTTGGGAGGAAG	CCGACTAGGCGAATTCACTTACTGACCGGCCTGGGCTGCTCTGAG	ATCGATCAGTCGATCATCATCGATCGATCGTCGTCGAG
SEU1	4: 128,651,525-74	TTATGTGGTTGGAGGTCGTTGGA	GCGCCTGCGCAGCTTTTCCGCACGCGCCTCGCCGGCGCGCGGCTCTCTCA	AGTCGTCGATCGTCAGTCGTCATCGATCGTCGTCGATCGTCGTCGTCA
SEU2	12: 91,348,696-745	GGTTGTTTTTTTTTTGGTTTGT	ACGCTCTTCGGTAGTAATCGCTTGTCCTTCGCACCTGGGTTGTCTCGCCC	GATCGTCGATAGTATCGTCGTCGATCATCGTGTCGTC

BUM1	1: 6,269,232-81	ATAAGAAGTTGTTGTTGTGTA	TCCGCTGCTTCCGCGACATGCAGAAGTGCGTAC AGGGGACGCGAGGGGAG	ATCGTCGATCGTCGATCATGTCAGAGTGTCGT ATCAGATCGTCGAGAG
BUM2	11: 35,441,865-914	ATAGATGGGTTTGTGTAAGGGAGA	GTTCGAGCCTTCCGGACGCCTTTGCATAAAAATG ACGAGACCTGTGCAGC	AGTCGAGTCGATCGATCGTCGATATGATCGAG ATCGTAGTCAGTC
BUU1	2: 38,177,672-708	TTATAGGGTATTTTATGGTTTAAG	GATGACGGCCTAGAAGGGAAGGGAGAGACTTAC TTTG	AGATGATCGATCGAGAGAGAGAGATCATCG
BUU2	10: 98,031,121-70	GAAAATTTTATTTTAATTTTTGA	CTGGCGGGGACGGTTATTTTATTAAGCTTGTCC ATTCTGTTTGGTAATTG	ATCGTCGAGTCGTATATAGTCGTCATCGTGTA TG
AMP1404	11: 118,213,647-96	ATTTATATATAATTTAAATATTGTTAT	ATCTCGAAGCCCCTTGAGAGAAGCCGTCGGCCC CATAGCGCAAGCCGTAG	GATCGAGTCTGAGAGTAGTCGTCGTCATAGTC GTCAGTCGTAG
AMP1730	9: 130,911,551-86 9: 130,911,627-42	TTGTTTTTGTTAGAGGTGTAGT TTAGGTGTAGAAATTTTGTTAAGT	ACTCCGGGAATGTCCCTCACTCTCCCCGTCCCTC TG GTTTCCGCAGGAGTTG	GATCGACTGTCTATCGTCG AGTCGTCAGAGTCG
AMP1746	11: 3,121,421-81	GTAGGGTATGGTAAGGGTGT	AATCCTCGGCTCGGTTCAGGAAGGTGAGCGTGG CTTTGCCGTCCAGCAGCGCCGACAGCGA	GATCGTCGTCAGAGTGAGTCGTAGTCGTCGTC AGTCAGTCGTCGATCAGTCGA
AMP1817	11: 66,083,571-611 11: 66,083,638-98	TTTGGTATTTATTTGTGTTTATAT TAGTGGTAGTTGTAGTTTTGTGGTTT	AGCGGTGCGGATCATCCTCCGCTGGCCGGAAAC CCAGGCGA ACCGGGCTCACACTGCTGCTCGCACGGAGCCTG GGCACAGGGGTCCTCGCAACTGCGCCCG	GAGTCGTGTCGATCAGTCGTCGTCGATCAGTCG GA GATCGTCATCATCGTCAGTCGTCATCGAGTCG TCATCAGTCGTCATCGTCGTCG
AMP2004	12: 6,900,126-84	TGTTTTTGTGGTGGATATTTTTA ATGTTGTAGGTTGATATAAAATGG	TCGGGGCAAGGCTAA CCGCCGCCCTCAAAGTCAGATGAAAGAGCCCCT GAGGACAGCGTTAGAGACACTCGGGA	GTCGATCAGTCTA ATCGTCGTCAGTCAGATGCAGAGTCGAGATCA GTCGTAGAGATCATCGA
AMP2007	2: 87,048,520-59 2: 87,048,581-638	TAAAGGAGTTGTTAATATTT TATAATTTTTAGTATAGTGTT	ACCTCGTTCTGCGGTTAAGAAACCAACAGGAAA AAGAACG GGCGCCTGTGAGGCACTCAGCCGACGGGAGCTT TGTTCTTCGTTGTATTGTGGCGGGG	GATCGTCGTCGATAGATCATCAGAGATCG AGTCGTCGTGAGTCATCAGTCGATCGAGTCAG TCGTGTATGTGTCG

[177]

5.2.6 Sample analysis

Genomic DNA was extracted using either the QIAamp DNA Investigator kit or the BioRobotEZ1® DNA extraction as described in sections 2.2.1.2 and 2.2.1.3. Then, to assess the quality and quantity of the resulting solution, DNA samples were quantified using the Quantifiler Human DNA investigation kit as described in 2.2.2.1. As previously mentioned in section 2.2.3, a total of three different bisulphite conversion kits were employed and their efficiency was evaluated through the built-in conversion controls. Bisulphite-converted DNA samples were amplified using the optimised PCR conditions for each assay together with pre-defined DNA methylation controls (0%-100%). Afterwards, 10 µl of biotinylated PCR products were sequenced on a pyrosequencer as described in section 2.2.6.1 and DNA methylation values were obtained via the PyroMark CpG software (QIAGEN). Specific experimental conditions including sample volumes will be mentioned in each results section separately.

5.3 Results

5.3.1 Evaluation of a Pyrosequencing®-based method (HBA1)

As discussed in the introduction, Pyrosequencing® was thought to overcome existing methods' limitations since it allows for absolute quantification of methylation levels. However, it was important to evaluate a Pyrosequencing®-based method using more stringent forensically recognised validation criteria. For the purpose of the initial Pyrosequencing® protocol evaluation, it was decided that a commercially available CpG assay for *HBA1* gene would be employed to ensure high PCR efficiency. Since HBA1 has been proposed as a blood-specific mRNA marker in the literature (Haas *et al.*, 2011b), it was also interesting to investigate the methylation status of this genomic locus. The efficiency of bisulphite conversion, the accuracy and linearity of methylation quantification as well as the overall sensitivity of the method were evaluated.

5.3.1.1 Bisulphite conversion rates

Apart from the four CpG sites, the assay also includes a built-in bisulphite conversion control [Figure 2-6]. To evaluate the efficiency of bisulphite treatment, 400 ng of six blood DNA samples were treated with Epitect® Bisulphite kit as described in section 2.2.3.1 and analysed using the proposed method. The average obtained peak height was 269±178 rfu while the average bisulphite conversion rate was calculated >99% in all samples, which was expected since 400 ng of starting DNA material was within the kit's optimal range. Since forensic specimens are likely to be of lower quantity, one of the blood samples was chosen to test decreasing amounts of DNA to be treated. Therefore, 10, 5, 2 and 1 ng of blood DNA were converted using the same experimental conditions, except the maximum DNA input into the PCR was used (7.5 μl). As a result, the obtained peak heights were significantly lower (average of 55 rfu); however, the conversion rate when using only 1 ng was still very high (96.6%). In all cases, blood samples showed low methylation levels (<0.08) with an average methylation value of 0.033. Since *HBA1* is expressed in blood, it was thought that the detected low methylation could indicate expression of this gene.

5.3.1.2 Sensitivity

To assess how decreasing starting material affects the methylation quantification, a blood sample was treated from 1 ng to 100 pg. As shown in Figure 5-4, the 'expected' methylation levels (<0.1) were obtained down to the smallest DNA input, although average methylation levels slightly increased when decreasing the starting DNA material. It is believed that this could be due to the decreasing observed peak heights, thereby increasing the background-to-noise ratio.

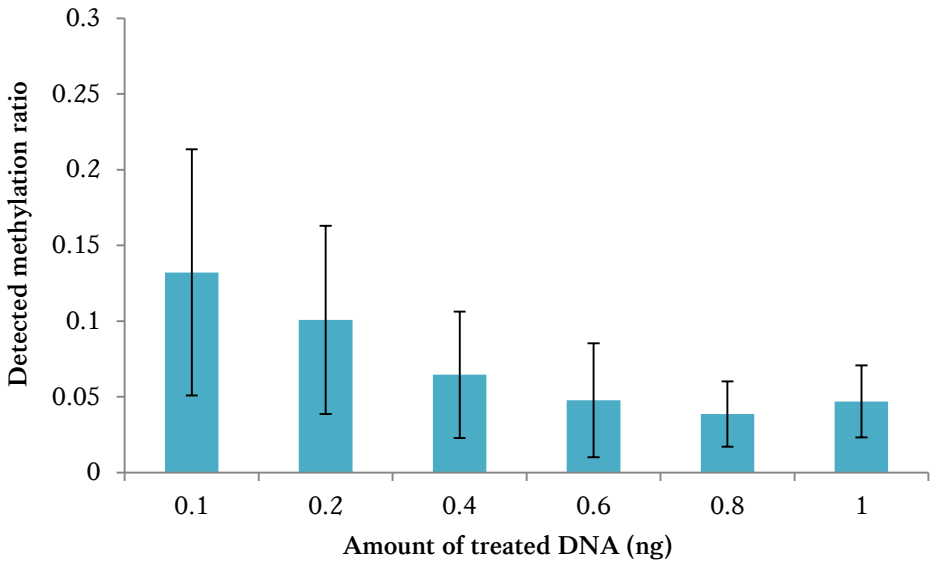


Figure 5-4. Average detected methylation levels with decreasing starting DNA amount using a single blood sample
Error bars represent standard deviation.

To further test this, 100 pg of a total of nine blood samples were analysed. As a result, the mean peak heights for all dispensations were >60 rfu while the obtained bisulphite conversion rates were >96% in all samples. As shown in Table 5-6, CpG sites were found to be unmethylated in most samples, but occasionally methylation was slightly higher. For example, for blood 8 CpG 4 was found to be completely methylated (>0.992). To assess if this was due to stochastic events because of the low template or due to natural inter-individual variability, 400 ng of this sample was reanalysed resulting in <0.1 methylation. Therefore, it can be concluded that caution is needed when analysing less than 1 ng of DNA since methylation levels could be altered due to either stochastic events during PCR or incomplete bisulphite conversion for particular CpG sites (the bisulphite conversion optimal DNA input range 200 to 500ng).

Table 5-6. Detected methylation levels using 100 pg of blood DNA (n=9)

HBA1 assay	Methylation ratio								
	Blood 1	Blood 2	Blood 3	Blood 4	Blood 5	Blood 6	Blood 7	Blood 8	Blood 9
CpG 1	0.090	0.040	0.091	0.022	0.311	0.019	0.034	0.072	0.022
CpG 2	0.040	0.026	0.061	0.013	0.035	0.013	0.023	0.036	0.020
CpG 3	0.026	0.011	0.085	0.006	0.014	0.012	0.031	0.032	0.017
CpG 4	0.034	0.010	0.067	0.014	0.006	0.000	0.026	0.992	0.005
Average	0.047	0.022	0.076	0.014	0.091	0.011	0.029	0.283	0.016

5.3.1.3 Reproducibility

To assess the accuracy of methylation quantification, a blood sample was analysed in six technical replicates. 100 ng of DNA was treated and 3 µl of converted DNA was used as a template in PCR. The obtained results were promising since for each CpG site the maximum standard deviation was 0.05.

5.3.1.4 Linearity

All blood samples analysed so far showed very low methylation levels; therefore it was important to assess how the method performs when analysing samples with partial methylation levels. Published research has shown that amplification bias towards either the methylated or the unmethylated allele is common in bisulphite PCR (Moskalev *et al.*, 2011). Even though extensive optimisation could eliminate possible bias, analysing pre-defined DNA methylation controls could be employed to 'correct' it. Commercially available low- and high- methylated controls were mixed in ratios before PCR analysis to create 10%, 20%, 30%, 50%, 70%, and 90% controls. Methylation levels on a DNA strand are linked with methylation levels in trans on the other strand for a particular CpG site, but also it is believed that methylation levels at a CpG site can very likely indicate similar methylation levels at cis-CpGs. Figure 5-5 shows the generated standard curve showing observed *vs.* expected methylation ratios. Interestingly, a level of at least 80% methylation was detected for all >50% controls indicating that the methylated allele of *HBA1* gene was undergoing significant preferential amplification. The introduced bias could be explained if one of the PCR primers contained a CpG site, which somehow favoured primer binding when it is methylated. However, this hypothesis could not be confirmed since the employed primer sequences were proprietary. The standard curve fitted a cubic polynomial line (R^2=0.975) that could be used to correct the observed non-linear quantification.

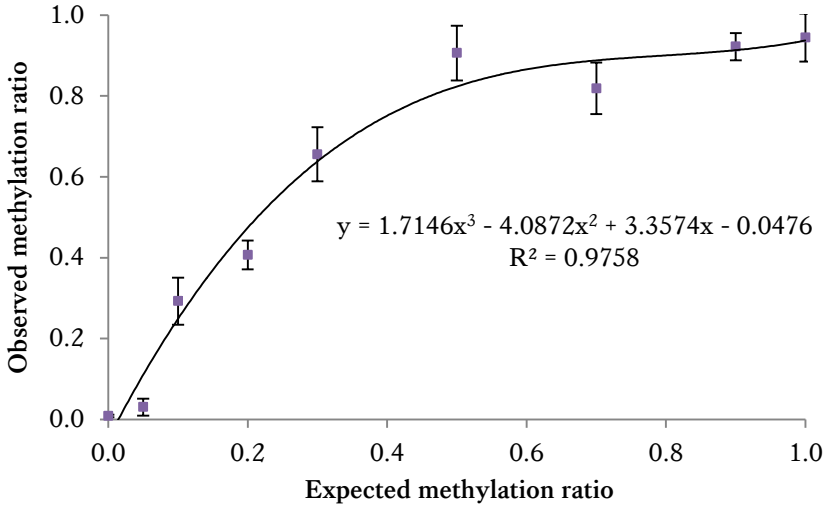


Figure 5-5. Observed vs. expected mean methylation ratio of pre-defined DNA methylation controls for HBA1 assay

5.3.1.5 HBA1 specificity

To evaluate the usefulness of the proposed assay in the identification of body fluids, initially five samples of each body fluid including blood, semen and saliva were analysed. 1 ng of each sample was bisulphite-treated and their methylation levels were obtained as before. Interestingly, all body fluid samples analysed showed <0.1 methylation for all four CpG sites indicating that this marker cannot be used for differentiating between the tested body fluids.

To sum up, taking into account the validation results using the pre-designed HBA1 assay, bisulphite Pyrosequencing® seems to be very promising for accurate methylation quantification of forensic samples. Initial results showed that the proposed method demonstrates high conversion rates (>96%), sensitivity (down to 100 pg) and accuracy (<0.05 standard deviation); however, caution is needed when minute DNA samples are analysed. Also, amplification bias could be common and alter detected methylation levels; therefore, analysis of pre-defined methylation controls is required to 'correct' biased methylation detection. Although pre-designed Pyrosequencing® assays could be very useful as extensive optimisation could be avoided, each assay analyses only a few CpG sites and identifying potential tissue-specific CpG sites could be a time-consuming process. Thus, selecting tissue-specific gene expression is not sufficient; it is important that the location of potentially useful CpG sites is identified before assay development.

5.3.2 Validation of a reported blood-specific marker (EFS)

5.3.2.1 Embryonal Fyn-associated substrate (EFS) gene

As mentioned before, Neumann *et al* (2011) detected differential methylation patterns of the *EFS* gene when analysing various human tissues including the forensically relevant tissues blood, sperm and buccal cells. *EFS* encodes for a protein that plays a role coordinating cell adhesion via tyrosine-kinase-based signalling. It has been recognised to be a member of the CRISPR-associated (CAS) protein family but little is known regarding its function. It contains a Src homology 3 (SH3) domain and has already been identified to interact with Src-family kinases in mice (Ishino *et al.*, 1997). Research has suggested that it is an adapter protein that is regulated via phophorylation but has no enzymatic activity (Tikhmyanova *et al.*, 2010). Additionally, its methylation mechanism indicates EFS involvement in the differentiation of the hematopoietic cell lineage (T-lymphocyte regulation) but has not been linked to cancer yet. However, other members of the CAS family are known oncogenes acting as prognostic markers of metastasis (Donlin *et al.*, 2005).

5.3.2.2 Optimisation of EFS assay

To avoid mis-priming, primer self-annealing or the generation of non-specific PCR products, the EFS assay was optimised using an annealing temperature gradient, various concentrations of $MgCl_2$ and primer, as well as different PCR cycling conditions. Figure 5-6 shows a critical step of the optimisation process where a set of six annealing temperatures (50-65 °C) were tested using a blood sample.

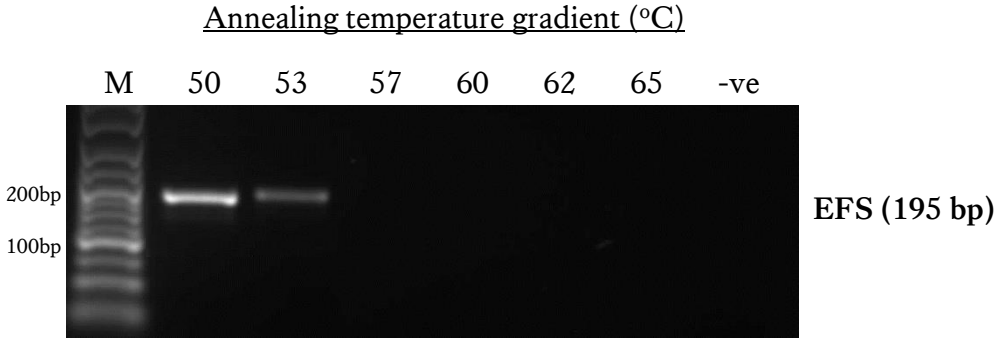


Figure 5-6. Final optimisation for EFS assay
Agarose gel image showing the results of the optimisation step regarding annealing temperature using a blood sample (expected PCR product length in brackets); the first column represents the DNA marker (M) while the next six columns show the resulted amplification bands for temperatures 50-65 °C. The last column represents the PCR negative (no-template) control.

5.3.2.3 Accuracy and linearity of methylation quantification

As shown when testing the linearity of HBA1 assay, non-linear methylation quantification was observed due to potential different amplification efficiencies of the unmethylated and methylated allele. To assess the performance of the EFS assay, 100 ng of each DNA methylation control (0-100%) (EpigenDx) were analysed in duplicate. The mean and standard deviation of every control and for each CpG site was calculated in order to assess the accuracy of the method. The average standard deviation obtained from the duplicate samples was 5%; however the maximum reached 18% (75% methylation control, CpG 8). Although the 5% could be considered as an acceptable degree of variability, a difference of 26% for the same CpG site of the same sample is regarded quite high. It is suspected to be due to events during amplification. In general, differences between CpG sites were observed, indicating the need to calculate the mean methylation per sample for more accurate results. Surprisingly, it was noted that the highly methylated control (100%) resulted in an average methylation of 74% (43-92%) [Figure 5-7a]. According to the manufacturer, this control demonstrates a genome-wide >85% methylation; however, variations in individual CpG sites cannot be excluded. Therefore, to build the standard curve of methylation quantification, the expected methylation values were 'normalised' using the values obtained for the non-methylated and methylated controls [Figure 5-7b].

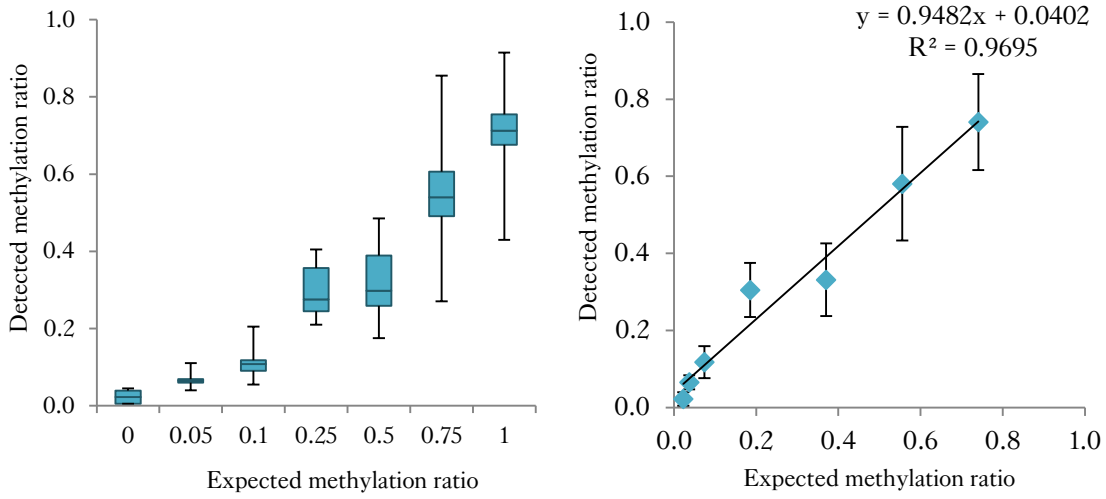


Figure 5-7. Linearity of methylation quantification for EFS assay
(a) Box plot showing the minimum, median and maximum methylation levels detected for each methylation standard taking into account all ten CpG sites, (b) Standard curve showing the observed *vs.* expected methylation ratio of the average methylation per standard after normalisation. Error bars correspond to standard deviation.

5.3.2.4 Verification of methylation patterns

Neumann *et al* (2011) had found that *EFS* is highly methylated in blood, completely unmethylated in sperm and partially methylated in buccal cells. To verify these patterns, 100 ng of one sample for these three tissues was analysed. Figure 5-8 shows that the confirmation of the original study's results was successful; taking into account only the first five CpGs the average blood methylation was 0.69, semen was completely unmethylated (0.01) while saliva resulted in partial methylation (0.44). However, especially for blood, there was a large variation among CpG sites (e.g. CpG 4 was 1, while CpG 5 was 0.46 methylated); therefore, more samples needed to be analysed to both confirm the specificity of EFS assay as well as investigate potential inter-individual variation in methylation levels.

5.3.2.5 EFS specificity

In order to apply *EFS* differential methylation patterns for the detection of blood, various forensically relevant tissues needed to be investigated; a total of 77 body fluid samples (20 blood, 10 semen, 16 saliva, 10 buccal, 10 vaginal fluid and 11 menstrual blood samples) were analysed. DNA samples were extracted as described in section 2.2.1.2 and the quantity of eluates was determined using the Quantifiler Human DNA quantification kit (section 2.2.2.1). The average DNA yield per µl of starting material was as follows: 17±7 ng for blood, 62±56 ng for semen and 14±31 ng for saliva. As for the swabs, a buccal swab yielded an average of 2,802±1,615 ng while a vaginal or menstrual secretion swab generated 2,247±1,542 ng and 2,657±2,487 ng of DNA respectively.

For each sample, 100 ng of DNA were bisulphite-treated and amplified using the proposed EFS PCR assay. All pyrograms™ passed the quality control of the instrument (expected sequence pattern, peak height >70 rfu) and the average bisulphite conversion rate of the eight built-in conversion controls was 94±5%. In general, it was noticed that the obtained conversion rates decreased slightly as the sequencing reaction progressed; possibly due to previously unincorporated cytosines. For each body fluid, box-and-whisker plots including data from the first and third quartile as well as the median (thin line) and minimum and maximum (error bars) methylation levels were calculated [Figure 5-9].

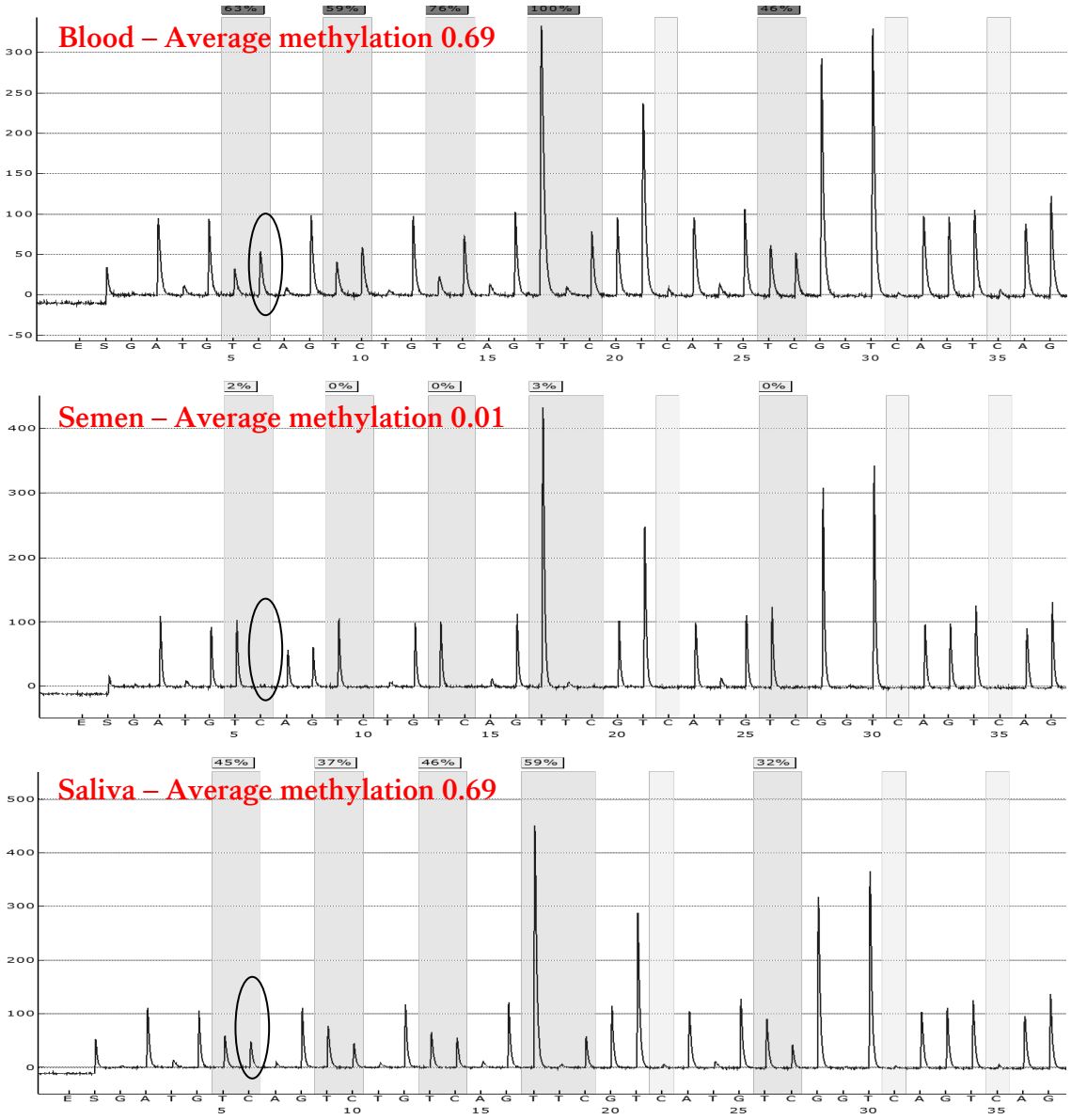


Figure 5-8. Verification of reported EFS methylation patterns in blood, sperm and saliva

As illustrated in Figure 5-9, the average methylation of blood was the highest among all body fluids (0.67±0.16), while semen was found completely unmethylated (0.06±0.04) in all samples except for two, where the mean methylation was 0.67 and 0.54 respectively (shown as outliers in Figure 5-9b]. It is believed that this could be due to either natural variability in methylation levels or possible contamination of semen with blood. The rest of tested tissues demonstrated partial methylation levels (saliva - 0.43±0.12, buccal cells - 0.26±0.1, vaginal fluid - 0.22±0.07 and menstrual blood - 0.22±0.05). Overall, there was a mean methylation range of 0.33±0.1 in every CpG site when taking into account all samples, with saliva showing the highest mean range of 0.5. This range corresponds to the observed inter-individual variability; therefore the lower it is the more robust the proposed CpG site is considered.

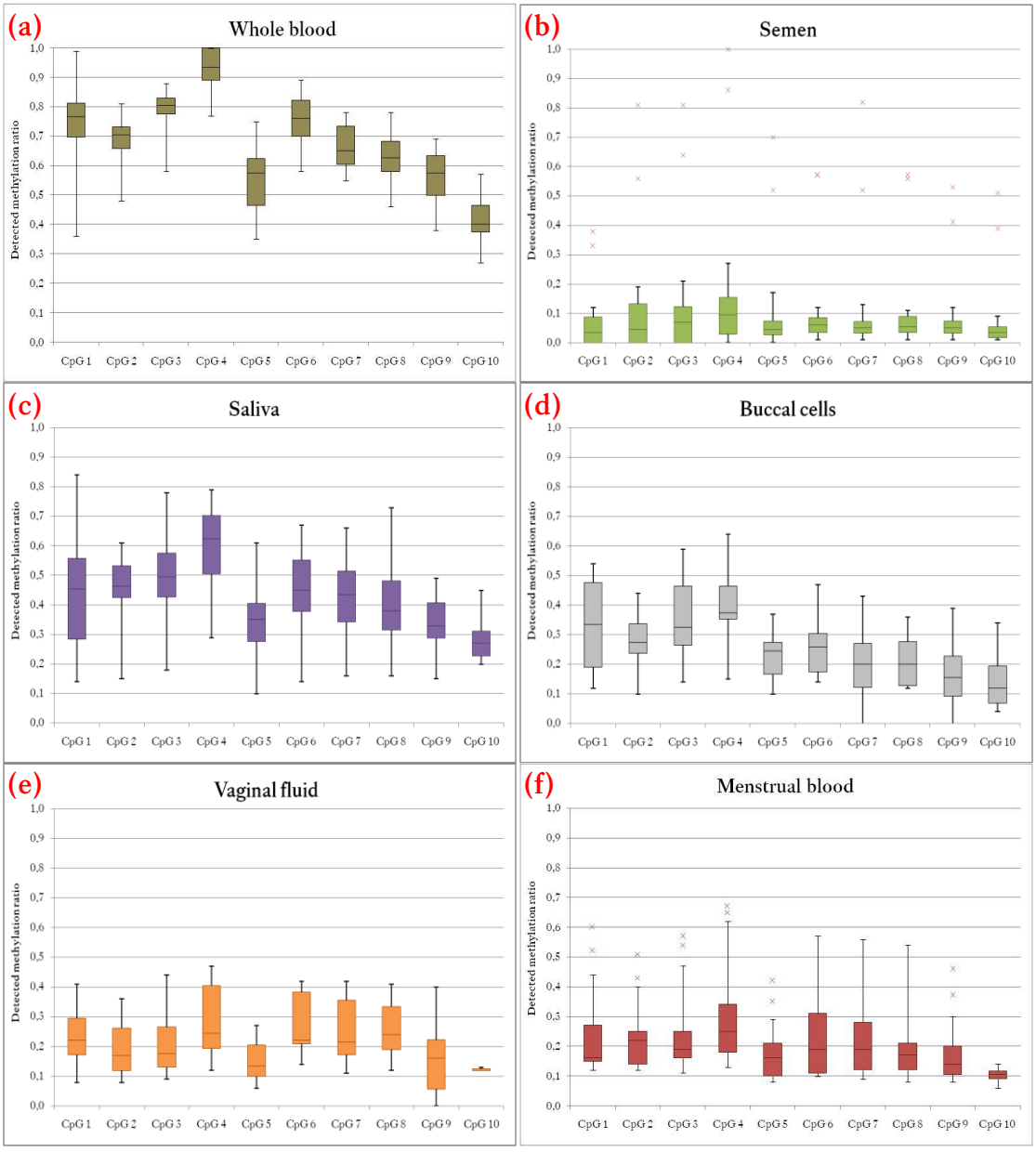


Figure 5-9. Box-and-whisker plots showing the detected methylation levels in (a) blood, (b) semen, (c) saliva, (d) buccal cells, (e) vaginal fluid, (f) menstrual blood for all ten CpG sites included in the EFS assay (total n=77)

Boxes represent the first and third quartile while the horizontal line represents the median value. Error bars correspond to the minimum and maximum detected methylation value. Outliers are shown as red 'x' dots.

Considering each CpG site separately, it was observed that the first four CpG sites (analysed by sequencing primer 1, Table 5-5), and especially CpG 4, seemed to be more promising as blood-specific differentially methylated markers. This is due to the fact that they show greater methylation difference between blood (~0.8) and the rest of the tissues (<0.6). Considering the limitations of existing methods including low sensitivity and specificity, a DNA methylation-based approach for the identification of blood using the proposed markers can be very beneficial either as a complementary

method or as a stand-alone assay (especially in cases where only DNA is available). However, caution is needed when analysing saliva as there is an overlap of methylation; a few saliva samples demonstrated for these four CpGs up to 0.8 methylation, more likely due to the presence of leucocytes in saliva.

5.3.2.6 Inter-individual variability of EFS methylation

To further test potential inter-individual variability of *EFS* methylation in blood, an independent cohort of blood samples was analysed. To account for possible age-, ethnicity- and gender-associated effects, this set included a total of 47 blood samples from female and male individuals aged 20-70 years from various ethnic backgrounds [Figure 5-10]. The exact experimental conditions employed in the specificity experiment were used (100 ng DNA) and only the first four CpG sites were analysed. The previously obtained levels [Figure 5-9a] were confirmed in this experiment with no significant methylation differences and no outliers.

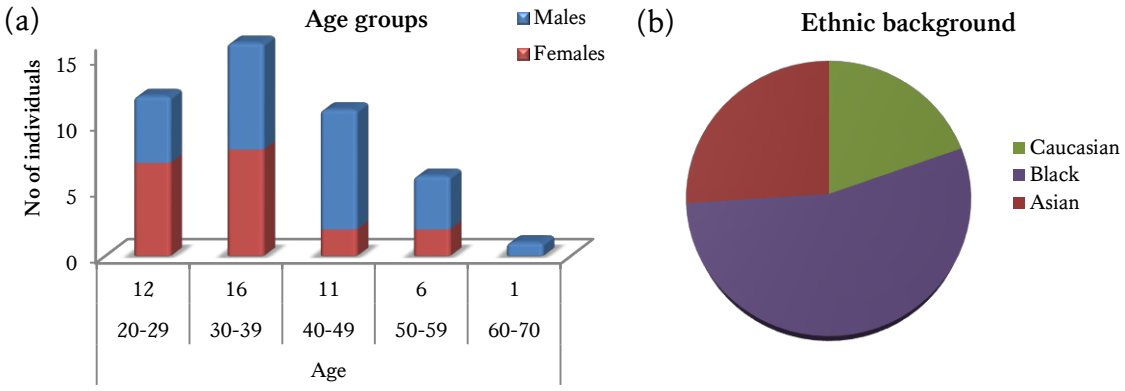


Figure 5-10. Information regarding (a) age and (b) ethnicity of the independent cohort of blood samples (n=47)

As shown in Figure 5-11, CpG 4 was confirmed to be the most useful marker since it not only demonstrates the highest level of methylation in blood (>0.8) but also the smallest variability. As shown in the graph, the median (horizontal line) is almost 1 (0.98). On the other hand, CpG 1 demonstrated a great variation amongst blood samples (0.2-0.9), which does not allow its use as a blood-specific marker. Lastly, there was no significant correlation between methylation and age, ethnicity or gender ($p>0.05$).

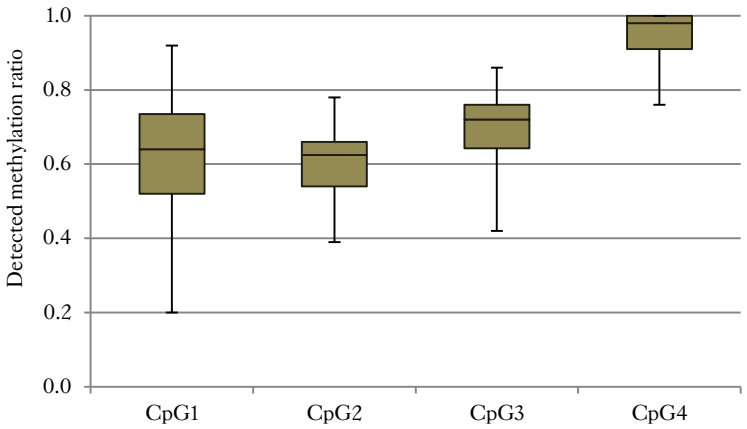


Figure 5-11. Inter-individual variation of methylation levels for the most blood-specific CpG sites of the EFS assay
Boxes represent the first and third quartile while the horizontal line represents the median value. Error bars correspond to the minimum and maximum detected methylation value.

5.3.2.7 Reproducibility of methylation quantification

To assess the reproducibility of quantification of the proposed Pyrosequencing® assay, 20 blood samples (10 ng DNA each) were bisulphite-treated and amplified in triplicate using the EFS assay. The mean and standard deviation of the observed methylation values were calculated for each sample and each CpG site as well. The mean standard deviation for CpGs 1, 2 and 3 was 0.08 while an average standard deviation of 0.05 was obtained for CpG 4 [Figure 5-12]. Although Pyrosequencing® has been proposed as a method that allows for accurate results, the increased number of PCR cycles (45) as well as the sequencing reaction itself could result in stochastic events. One can understand that choosing markers that have at least 40-50% difference between the tissue of interest and the other tissues is essential. Analysing in replicates could also serve as a solution since in some cases the observed standard deviation was as high as 26%.

5.3.2.8 Applicability in forensic casework

It is essential that the observed methylation status of the *EFS* gene when analysing freshly-collected body fluids is also detected in samples that are of low quality. To assess the stability of *EFS* methylation as well as its applicability as a blood-specific marker in forensic casework, a set of mock casework samples were prepared and analysed.

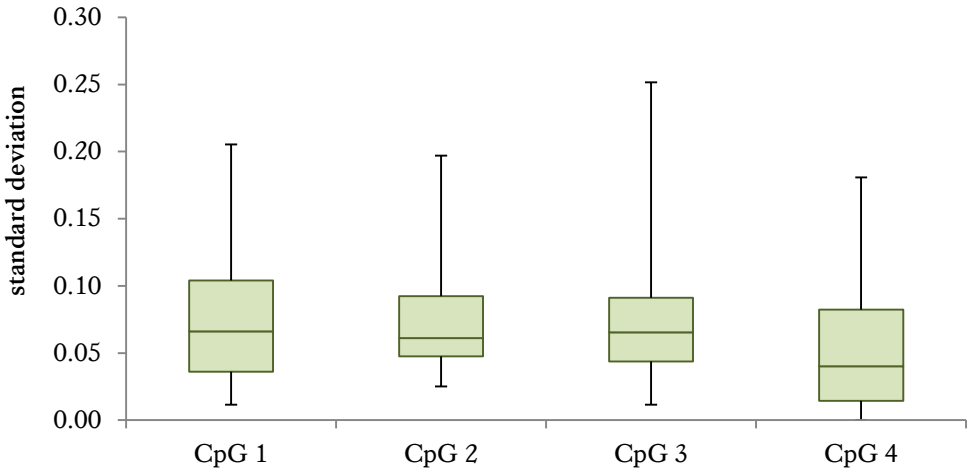


Figure 5-12. Standard deviation of methylation quantification of each CpG site as obtained when analysing 20 blood samples in triplicate
Boxes represent the first and third quartile while the horizontal line represents the median value. Error bars correspond to the minimum and maximum detected methylation value.

As mentioned in Chapter 3, one of the most significant advantages of applying DNA methylation profiling for the identification of tissues is its potential applicability in cold cases. In most of these cases, mRNA profiling cannot be employed either because in stored stains mRNA molecules degrade over time or because only DNA samples are stored. To assess the stability of *EFS* methylation, a total of five blood stains stored at room temperature protected from light for 9-18 years as well as five blood DNA samples stored at -20 °C were analysed. In all cases, the expected methylation ratios for the investigated CpG sites were obtained (CpG 1 – 0.72±0.05, CpG 2 – 0.63±0.05, CpG 3 – 0.72±0.05 and CpG 4 – 0.94±0.08).

Furthermore, depending on case circumstances forensic specimens could be mixed or degraded due to sun light, temperature and weather conditions. In an attempt to recreate 'forensic-like' scenarios, various stains were prepared as described in section 5.2.1.3 and their methylation level regarding the first four EFS CpG sites was quantified. The whole swab or stain was used for DNA extraction; then, 10 µl of extracted DNA was used for bisulphite treatment and converted DNA was eluted in 10 µl.

Briefly, all stains stored at various temperatures for a week yielded the 'expected' blood methylation pattern indicating that under these conditions the methylation of *EFS* gene is stable. This is very important as crime stains are often exposed to sunlight, so it is encouraging for DNA methylation-based applications to prove that methyl

groups are stable under these temperatures. On the other hand, regarding the artificially UV-degraded samples, the 'expected' methylation pattern was observed for all bloodstains incubated under UV for up to 90 minutes. The stain that was UV-degraded for 120 minutes resulted in a completely non-methylated profile. This could indicate either that the methylation levels were altered due to the UV light or that a non-blood tissue is present. Potential stochastic effects and the effect of drop-in contamination cannot also be excluded as it is known to be common in highly degraded stains. However, this potential contamination is expected to be very low-level as following STR profiling the obtained DNA profiles from all time-points were found to be single-source. The stain that was incubated for 4 hours under UV light was too degraded to produce an amplicon.

Additionally, bloodstains on the towel and tissue paper gave successful pyrograms™; however, the methylation obtained from the washed blood stain was very low (0.08). The stained shirt was washed in a public washing machine; therefore potential contamination with other body fluids/tissues or the effect of chemical reagents could explain this result. These factors should be taken into account when interpreting methylation profiles from crime scene stains. Also, a reduced methylation was obtained from the stain on jeans (average of 0.33); however in this case it could be either due to natural inter-individual methylation variation or due to the effect of dyes on the analysis. Lastly, blood, semen and saliva were used to generate two mixed stains; the obtained methylation values of both single-source and mixed stains regarding CpG 4 are presented in Figure 5-13. Since mixture analysis is considered one of the main drawbacks of DNA methylation-based tissue identification, it was important to assess how this assay performed when analysing mixed stains. As illustrated, using the observed methylation of single-source tissues, the 'expected' methylation value was calculated. Although in both cases the observed methylation ratio was not significantly different from the 'expected' value, correction via the equation of the linear regression line in the standard curve for CpG 4 [Figure 5-13b] improved the accuracy of the results. These results are very encouraging; however, the analysis of more mixed stains generating in various ratios is necessary prior to any conlcusions made.

(a)

Samples		CpG 4
Blood	Observed	0.97
Semen	Observed	0.00
Saliva	Observed	0.51
Blood:Semen (1:1)	Expected	0.49
	Observed	0.61
	Corrected	0.53
Blood:Saliva (2:1)	Expected	0.82
	Observed	0.91
	Corrected	0.84

(b)

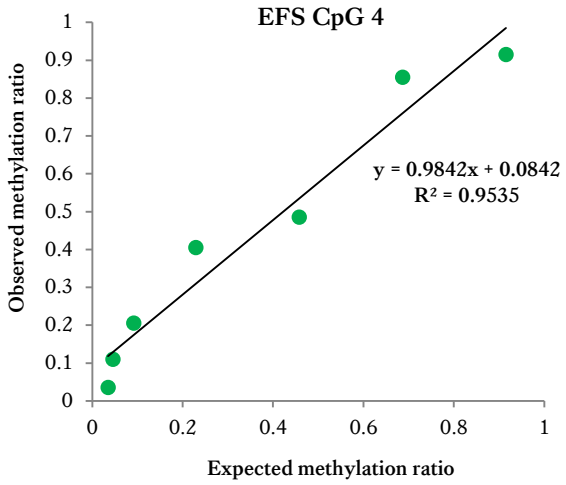


Figure 5-13. Methylation analysis of EFS CpG 4 in mixed stains

(a) Observed methylation for all individual and mixed samples. Individual methylation values were used to calculate the expected values for the mixed stains. (b) Standard curve for EFS CpG 4 showing the observed *vs*. expected methylation ratios of pre-defined DNA methylation controls. The equation of the fitted linear regression line was used to 'correct' the observed methylation for mixed stains.

5.3.3 Analysis of genome-wide DNA methylation data

5.3.3.1 Optimisation of Pyrosequencing®-based assays

As described in section 5.2.2.2, a total of twelve potentially tissue-specific CpG sites were identified. All markers belong to a gene and are found either in the 5'-end region or within the main body of the gene. A bisulphite Pyrosequencing® assay for each CpG was designed using the BiSearch software and following the guidelines mentioned in section 2.2.4.1. However, it was necessary to optimise the PCR reactions to avoid mis-priming, primer self-annealing or the formation of non-specific PCR products. Each assay was optimised using an annealing temperature gradient, various concentrations of $MgCl_2$ and primer, as well as different PCR cycling conditions. The optimisation of the assay designed for BLM2 marker proved to be very challenging as the amplification efficiency was very low in all tested PCR conditions; therefore, it was excluded from further analysis. Figure 5-14 shows the successful amplification of three samples using the optimised PCR conditions for each marker (11 in total).

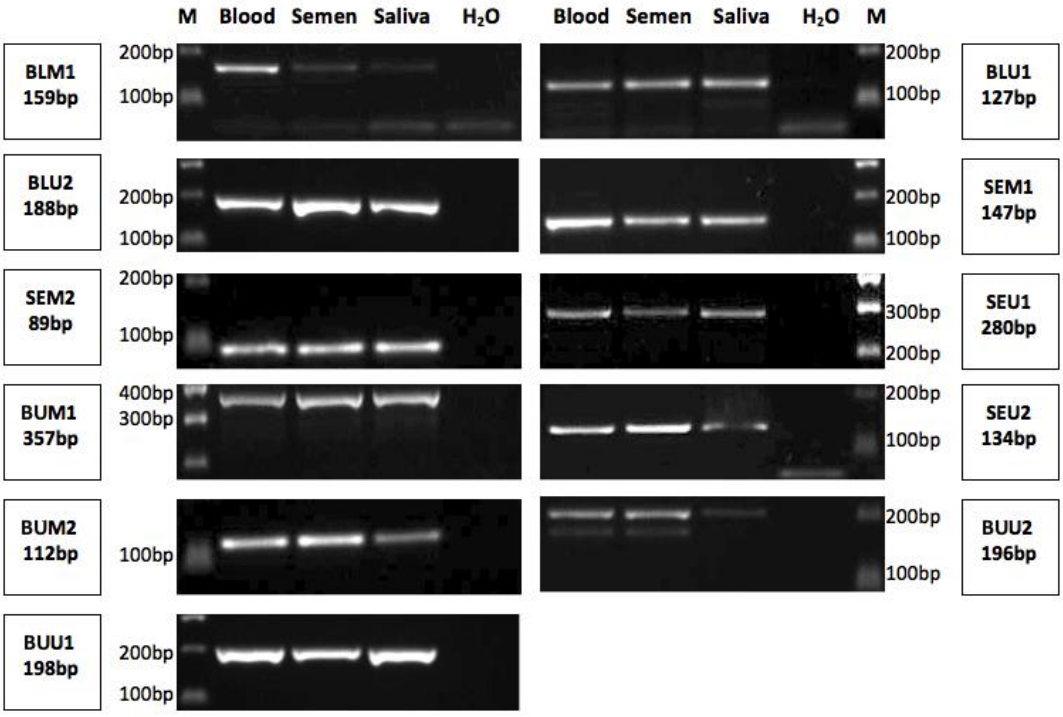


Figure 5-14. Final optimised PCR amplicons

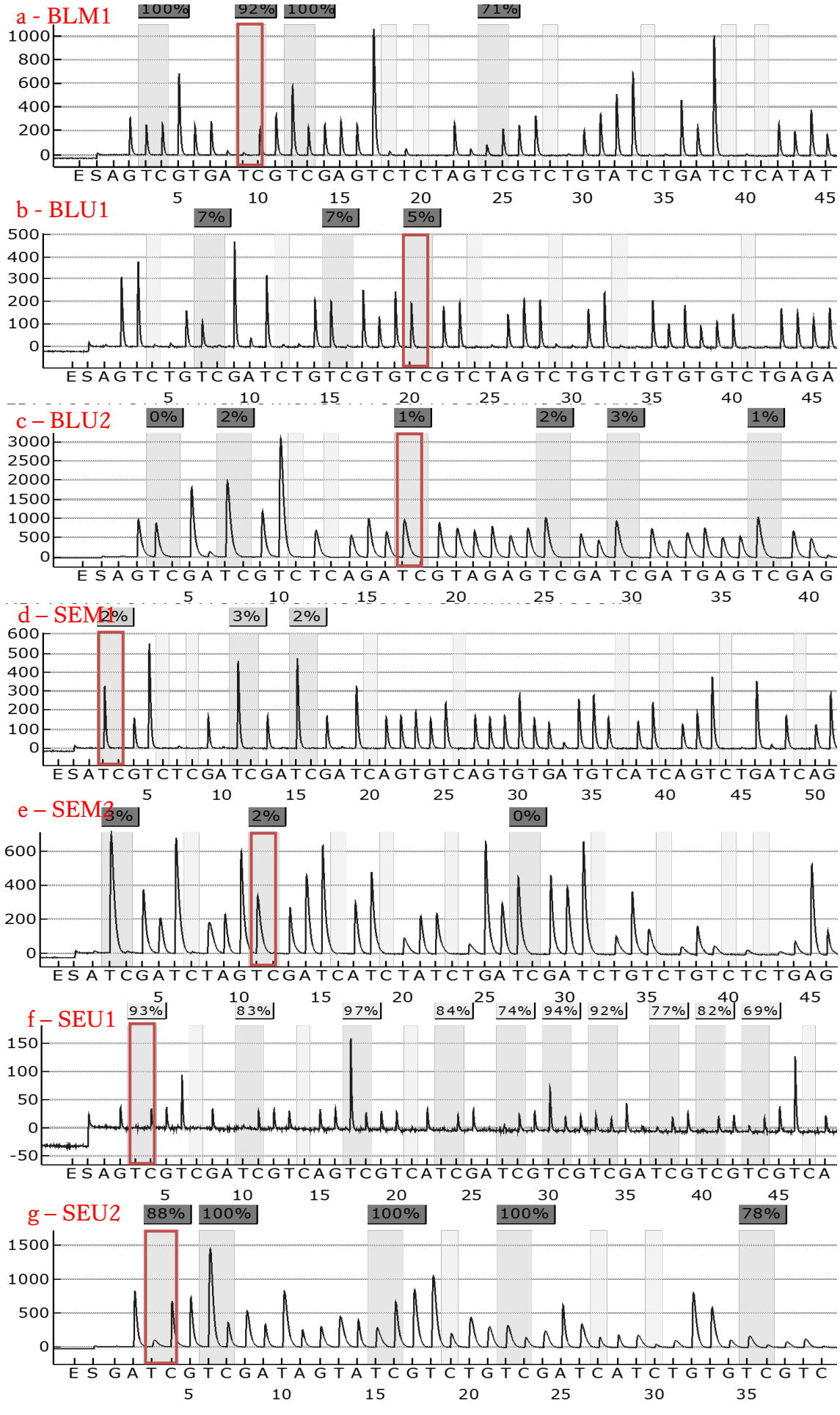
Successful amplification of three samples (blood, semen and saliva) using all eleven assays. BLM2 failed optimisation so it is not included here. Expected PCR lengths are shown in the boxes, together with the expected fragments of the DNA marker on the side of the gel images. As shown, for some markers (e.g. BLM1 and BUU2), there was a difference in the bands' intensity, however it is believed to be due to loading errors rather than differences in PCR efficiency among body fluids. M: DNA marker

Optimised PCR products were then sequenced and methylation quantification was performed. Figure 5-15 illustrates example pyrograms™ obtained from one blood sample after analysing for all designed assays. Bisulphite conversion rates were calculated using the peak heights of thymine and cytosine of each bisulphite conversion control (non-CpG site cytosines) and were in the majority of cases >98%. If bisulphite conversion rates were lower than 95%, the treatment with sodium bisulphite was repeated. The developed assays demonstrated different PCR efficiencies as indicated by the peak heights in Figure 5-13, with SEU1 showing the lowest peak heights. Additionally, for certain assays the peaks were too wide indicating too much PCR product input or in other cases, the signal decreased towards the end of the sequence probably due to incomplete incorporation of nucleotides [Figure 5-15g]. Lastly, there were cases where there was a low signal detected for internal 'dead dispensations' (where no signal is expected); however, it is believed that it was due to signal carry-over from previous dispensations rather than another fragment being sequenced.

5.3.3.2 Verification of methylation patterns

To verify the methylation patterns of the CpG sites reported by Rakyan and co-workers (2008 & 2010) to be specific for blood, semen and buccal cells [Table 5-2], a set of samples were analysed for each assay. It was important to confirm the methylation levels of the proposed CpG sites in these three types of tissues first, before any analysis is performed using other forensically relevant tissues such as menstrual blood or vaginal fluid. However, since the reason for choosing markers that are differentially methylated in buccal cells was their potential use in the identification of saliva, a set of saliva samples were co-analysed. The aim of this experiment was not only to verify the previously reported methylation but also potentially select the best CpG sites for further validation. Most assays also investigate adjacent CpG sites and their potential in identifying the tissue of origin was tested as well.

For this analysis, a total of 1 ng of each DNA sample was bisulphite treated using the EpiTect bisulphite kit (section 2.2.3.1) and bisulphite converted DNA was eluted in 20 µl of elution buffer. Then, 1 µl of converted DNA was used in each PCR reaction. Pyrosequencing® reactions were performed as described in section 2.2.6.1.


a - BLM1
100%
92%
100%
71%
E S A G T C G T G A T C G T C G A G T C T C T A G T C G T C T G T A T C T G A T C T C A T A T
b - BLU1
7%
7%
5%
E S A G T C T G T C G A T C T G T C G T G T C G T C T A G T C T G T C T G T G T G T C T G A G A
c – BLU2
0%
2%
1%
2%
3%
1%
E S A G T C G A T C G T C T C A G A T C G T A G A G T C G A T C G A T G A G T C G A G
d – SEM1
2%
3%
2%
E S A T C G T C T C G A T C G A T C G A T C A G T G T C A G T G T G A T G T C A T C A G T C T G A T C A G
e – SEM2
3%
2%
0%
E S A T C G A T C T A G T C G A T C A T C T A T C T G A T C G A T C T G T C T G T C T C T G A G
f – SEU1
93%
83%
97%
84%
74%
94%
92%
77%
82%
69%
E S A G T C G T C G A T C G T C A G T C G T C A T C G A T C G T C G T C G A T C G T C G T C G T C A
g – SEU2
88%
100%
100%
100%
78%
E S G A T C G T C G A T A G T A T C G T C T G T C G A T C A T C T G T G T C G T C

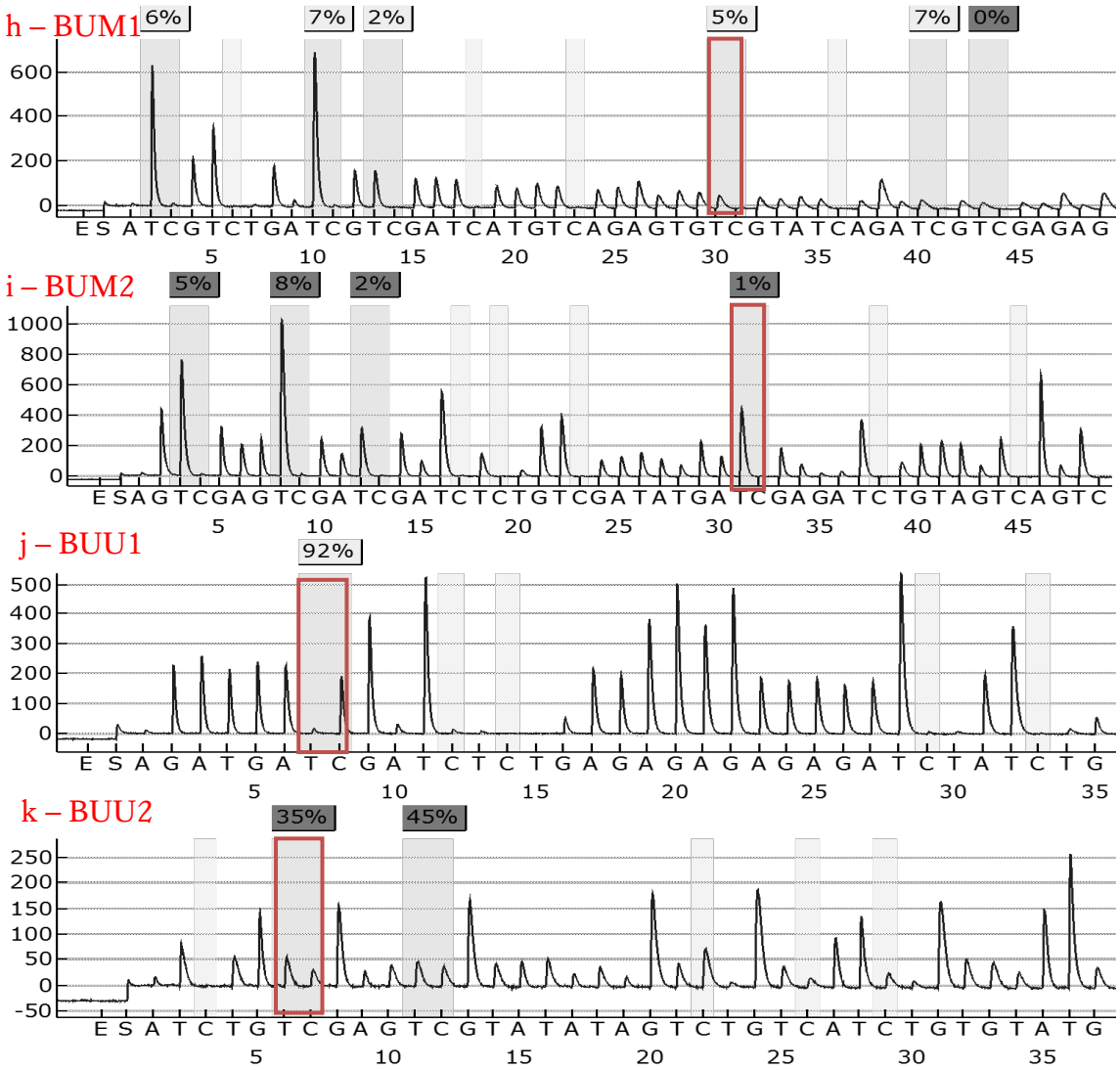


Figure 5-15. Example pyrograms™ of all designed Pyrosequencing® assays
Although there are eleven CpGs in question (highlighted in red blocks), in some assays adjacent CpGs were also quantified (dark grey boxes). Obtained methylation values are presented on the top, where the peak for 'T' corresponds to the unmethylated fraction (together with adjacent Ts or non-CpG converted Cs) while the peak for 'C' to the methylated fraction only. Light grey boxes indicate the position of built-in bisulphite conversion controls.

Buccal cell-specific markers

The markers cg15731815 (BUM1) and cg08258650 (BUM2) have been reported to be methylated in buccal cell samples (0.84 and 0.77 respectively) while demonstrating low methylation levels in blood and semen (0.05-0.08) [Table 5-2]. Although in the current study, their methylation status in blood and semen was confirmed, the reported methylation ratio in buccal cells was much lower (average of 0.47 and 0.48 accordingly) [Figure 5-16b]. Also, even though it was believed that saliva mainly consists of buccal cells in the oral fluid, therefore buccal cells and saliva samples would

share the same methylation profile, this was not observed using the proposed markers. Saliva samples resulted in methylation ranging from 0-0.4. Since the mean methylation value for saliva samples were 0.17 and 0.12 respectively, these markers were excluded from further analysis as the methylation difference with blood and semen was too narrow.

(a)

Assay	CpG site	Number of samples			
		Whole blood	Semen	Buccal cells	Saliva
BUM1	**cg15731815**	14	12	8	4
BUM2	**cg08258650**	15	11	7	12
BUU1	**cg05761971**	14	10	8	15
BUU2	**cg16779976**	11	9	9	10

(b)

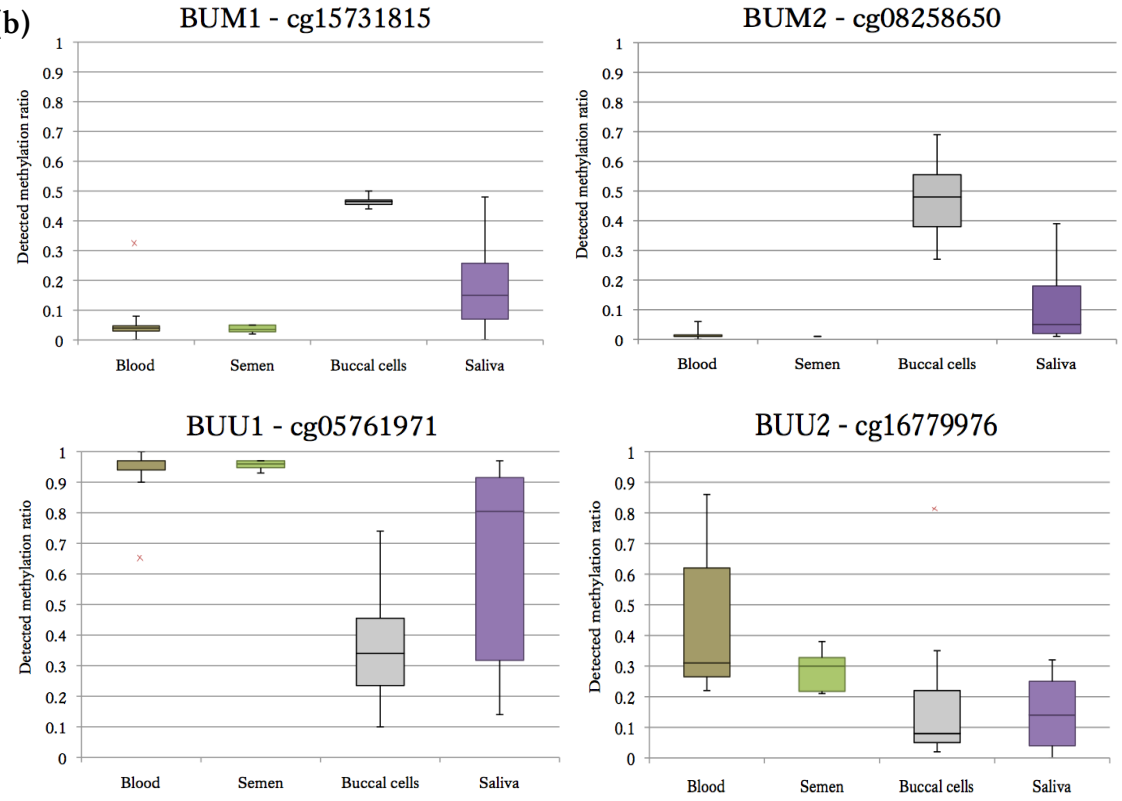


Figure 5-16. Methylation levels of the proposed buccal cell-specific markers in blood, semen, buccal cells and saliva

(a) Number of samples used in the verification experiment, (b) Observed methylation ratio of each of the CpG sites of interest included in four assays (BUM1, BUM2, BUU1, BUU2). Data are presented in the form of box-and-whisker plots showing the first and third quartiles (boxes), the median (horizontal line) and minimum and maximum (error bars) methylation values detected. Outliers (methylation value ≥3*SD) are shown separately as red 'x' dots.

The marker cg05761971 (BUU1) has been previously reported to be unmethylated in buccal cells (0.108), while being highly methylated in blood and semen (0.87 and 0.88 respectively). Similar to the previous two CpG sites, the methylation difference in this study was smaller than originally obtained as buccal cells gave an average methylation ratio of 0.37. The marker also demonstrated a large inter-individual methylation variation in saliva (n=15), hence it was concluded that it is not suitable for saliva identification. Finally, although cg16779976 was reported to be methylated in blood and semen by the genome-wide analysis, this observation was not confirmed by bisulphite Pyrosequencing®. The methylation of buccal cells and saliva was lower than the other two body fluids and quite similar but not discriminatory enough to include the marker in further analysis.

In conclusion, although the methylation patterns of three out of four selected CpG sites were verified in blood and semen, the methylation difference of these body fluids and buccal cells samples was less than previously measured [Table 5-2]. Additionally, saliva did not share the same methylation profile with buccal cells but especially for cg15731815 (BUM1) and cg08258650 (BUM2), gave similar methylation values with blood. It is believed that this could be due to the presence of leucocytes in saliva (e.g. gum bleeding). The amount of 'contaminating' blood cells or other cell types apart from buccal cells in saliva could vary between individuals and this could potentially explain the large inter-individual methylation ratio observed in some cases.

Blood-specific markers

Although there were a total of four blood-specific CpG sites selected following the analysis of genome-wide methylation data, only three of them were tested here since one assay (BLM2) failed optimisation. The number of samples analysed per assay ranged from 55-66 including 10-20 samples per tissue [Figure 5-17a]. The CpG sites cg17518965 (BLU1) and cg26285698 (BLU2) were reported to be non-methylated in blood (0.02 and 0.09 respectively) while being highly methylated in semen and buccal cells (0.86-0.96). In this study, only blood's methylation profile was confirmed, whereas semen and buccal cells demonstrated large inter-individual variation in detected methylation [Figure 5-17b].

Interestingly, there were two semen samples that had very low methylation (outliers), either because of natural variation in methylation levels and/or possible presence of blood in semen. Previous research has reported the presence of white blood cells in semen samples of infertile men or men with bacterial infections (Lackner *et al.*, 2006). Moreover, these particular two semen samples resulted in a 10-fold decrease in DNA yield following DNA isolation using the same starting material as the rest of the semen samples, potentially indicating a smaller number of spermatozoa present. Lower sperm count is often associated with infertility problems and it has been shown that alternations in sperm DNA methylation at particular loci are common with low sperm motility and different types of male infertility (El Hajj *et al.*, 2011; Hammoud *et al.*, 2010; Pacheco *et al.*, 2011). Furthermore, saliva samples once again demonstrated very low methylation, even though it was originally thought that they would be methylated at these two CpG sites. Even though, it was thought that cg17518965 could act as a semen-specific marker, it was decided that the similarity of methylation profiles between semen and buccal cell samples would not be ideal from a forensic perspective. Since buccal cells are of epithelial tissue origin, it came as no surprise when vaginal and skin samples resulted in similar methylation values with buccal cells when tested with this assay (data not shown).

(a)

Assay	CpG site	Number of samples			
		Whole blood	Semen	Buccal cells	Saliva
BLU1	cg17518965	15	12	9	19
BLU2	cg26285698	15	12	10	20

(b)

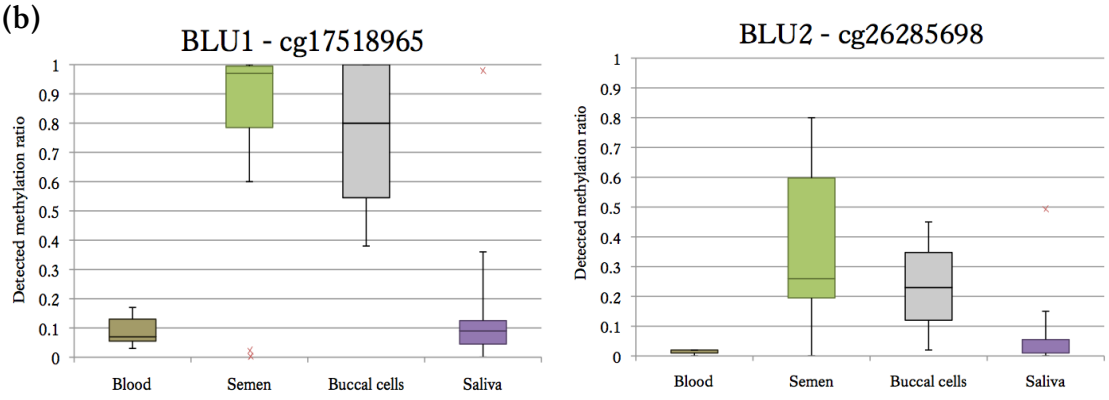


Figure 5-17. Methylation levels of cg17518965 and cg26285698 in blood, semen, buccal cells and saliva

(a) Number of samples used in the verification experiment, (b) Observed methylation ratio of each of the CpG sites of interest included in two assays (BLU1, BLU2). Data are presented in the form of box-and-whisker plots showing the first and third quartiles (boxes), the median (horizontal line) and minimum and maximum (error bars) methylation values detected. Outliers (methylation value ≥3*SD) are shown separately as red 'x' dots.

On the other hand, findings regarding cg13763232 (BLM1) were more promising. Initial results analysing only blood, semen, buccal cells and saliva samples revealed a distinct blood methylation profile as all blood samples resulted in >0.85 methylation ratio. Semen and buccal cells verified the expected low methylation since they had an average of 0.12 and 0.21 respectively. Also, even though the methylation levels of saliva samples ranged between 0.16-0.87 (mean=0.60), it was decided that the analysis of other forensically relevant tissues was needed before final conclusions regarding this marker's blood specificity are made. Thus, an additional set of 34 samples (9 vaginal fluid, 14 menstrual blood, 5 skin and 6 urine samples) were analysed; detected methylation levels are shown in Figure 5-18.

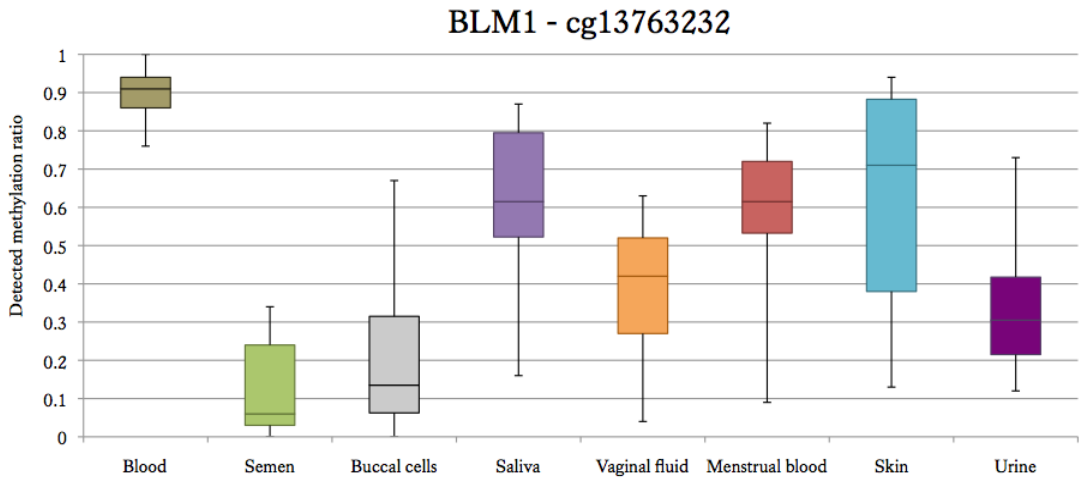


Figure 5-18. Methylation levels of cg13763232 in various body fluids/tissues
Observed methylation ratio of the CpG sites of interest (BLM1 assay) in various body fluids/tissues. Data are presented in the form of box-and-whisker plots showing the first and third quartiles (boxes), the median (horizontal line) and minimum and maximum (error bars) methylation values detected.

As shown in Figure 5-18, cg13763232 was shown to be highly methylated (>0.85) only in blood, while the rest of the tested tissues demonstrated various profiles being either non- or partially methylated. Similar methylation patterns were also obtained for the three adjacent CpG sites, indicating that it is not only cg13763232 that demonstrates blood-specific methylation but also the locus itself.

Furthermore, even though only five skin samples were analysed, a wide range in methylation ratio (0.13-0.94) was observed. From a forensic perspective, in cases that a stain is to be analysed and DNA methylation profiling is used to confirm positive results of presumptive tests, then the fact that some skin samples are highly methylated

should not be an issue due to the stain's colour and characteristics. However, in case of an invisible stain or when a DNA sample rather than a stain needs to be tested, employing the BLM1 assay would not be able to confirm the presence of blood with confidence. On the other hand, differentiation between blood and menstrual blood is often required in rape cases where a female in menses is involved. As shown in Figure 5-18, the methylation ratios obtained from whole blood and menstrual blood samples hardly overlap meaning that this marker could be used not only for the identification of blood, but also for excluding the presence of menstrual blood if needed. However, although there were only two menstrual blood samples (out of 15) that resulted in >0.8 methylation, analysing more blood and menstrual blood samples would reveal potential inter-individual variations.

Semen-specific markers

For the identification of semen, a total of four CpG sites were chosen; two (cg04382920 and cg11768416) were reported to have no methylation in semen (<0.02), while the other two (cg01318557 and cg05656364) were more than 0.9 methylated in Rakyan and co-workers' studies (2008 & 2010). The exact opposite methylation status was seen in blood and buccal cells [Table 5-2]. These patterns were confirmed when initially analysing samples of these three tissues (blood, semen and buccal cells) indicating their potential in the identification of semen. However, more tissue types needed to be investigated in order to establish their promising specificity. As a result up to 110 samples including various tissue types were analysed per assay [Figure 5-19a].

Regarding cg01318557 (SEM1) and cg05656364 (SEM2), all non-semen tissues including whole blood, buccal cells, saliva, vaginal fluid, menstrual blood, skin and urine demonstrated very low levels of methylation (mean methylation range of 0.01-0.06 for cg01318557 and 0-0.06 for cg05656364). In contrast with most of the potential blood- and buccal cell-specific CpG sites tested so far, buccal cells and saliva samples showed identical methylation ratios. On the other hand, semen samples resulted in a mean of 0.40 and 0.59 for cg01318557 and cg05656364 respectively; although there was a significant methylation range among samples [Figure 5-19b].

(a)

Assay	CpG site	Number of samples							
		Blood	Semen	Buccal cells	Saliva	Vaginal fluid	Menstrual blood	Skin	Urine
SEM1	cg01318557	10	12	15	12	10	15	9	7
SEM2	cg05656364	15	14	15	20	10	15	10	10
SEU1	cg04382920	9	9	12	11	10	10	2	5
SEU2	cg11768416	15	13	15	20	10	15	10	10

(b)

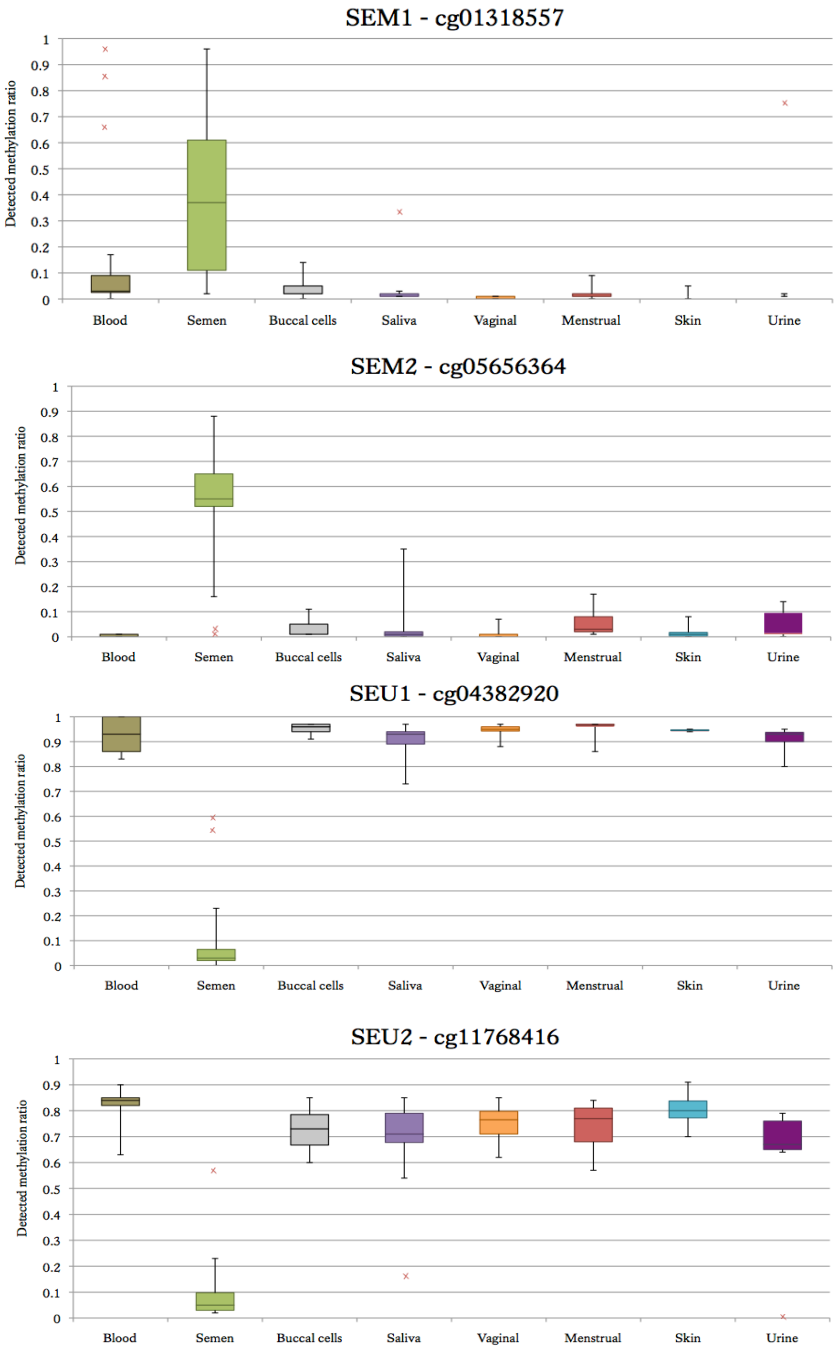


Figure 5-19. Methylation levels of the proposed semen-specific markers in various forensically relevant body fluids/tissues

(a) Number of samples used in the verification experiment, (b) Observed methylation ratio of each of the CpG sites of interest included in four assays (SEM1, SEM2, SEU1, SEU2) in various body fluids/tissues. Data are presented in the form of box-and-whisker plots showing the first and third quartiles (boxes), the median (horizontal line) and minimum and maximum (error bars) methylation values detected. Outliers (methylation value ≥3*SD) are shown separately as red 'x' dots.

Occasionally, there were a few outliers obtained for blood, saliva and urine in cg01318557 (SEM1), which is believed to be due to natural methylation variability or potential contamination of semen in urine since the particular sample belonged to a male volunteer. Although semen clearly demonstrated a different distribution in methylation compared to other tissues, it was believed that the observed outliers could introduce uncertainties in confirming semen using this CpG site. Similarly, for cg05656364 (SEM2) there were two saliva samples resulting in 0.32 and 0.35 methylation ratio respectively, while there were two semen samples that were found completely unmethylated even though this CpG site was proposed to be methylated in sperm. Interestingly, these two semen samples were the same with the ones showing an 'opposite' methylation profile in the BLU1 assay and were the ones with low sperm count. Since these samples represent 15% of the total analysed semen samples, it was thought that this marker will not be included in further analysis in this study.

On the other hand, as illustrated in Figure 5-19b, both cg04382920 (SEU1) and cg11768416 (SEU2) seemed to be highly specific markers for semen. The results of the genome-wide methylation analysis were confirmed for both markers, although the obtained methylation for cg11768416 in non-semen tissues was slightly lower than originally reported (0.75). Even though there were one or two semen samples per assay demonstrating higher methylation levels, these were not as high as the ones obtained by the non-semen samples. Additionally, there was one saliva sample (out of 20) that showed low methylation levels, but this is believed to be due to natural methylation variability. Thus, using these two CpG sites (cg04382920 and cg11768416) no false negative results were obtained and only one out of the total 154 non-semen samples resulted in a false positive identification of semen, which is particularly important in a forensic scenario. Interestingly, most of the co-analysed adjacent CpG sites (nine for SEU1 and four for SEU2) also demonstrated the observed semen-specific methylation pattern; therefore they can be used together with the two proposed CpG site to strengthen the identification of semen.

5.3.3.3 Validation of SEU1 and SEU2 assays

Without a doubt, CpGs cg04382920 (SEU1) and cg11768416 (SEU2) demonstrated a semen-specific DNA methylation profile; however, in order to implement such markers in forensic casework extensive validation of the associated methylation assays

is required. Initial validation of these markers included sensitivity and methylation quantification linearity analysis as well as testing aged semen samples.

5.3.3.3.1 Linearity of methylation quantification

As shown in previous bisulphite PCR assays [Figure 5-5], non-linear methylation quantification was observed due to potential different amplification efficiencies of the unmethylated and methylated allele. To assess the accuracy of SEU1 and SEU2 assays, 100 ng of each DNA methylation control (0-100%) (EpigenDx) were analysed in duplicate. The mean and standard deviation of each standard was then calculated taking into account all CpG sites included in the sequences (ten CpGs in SEU1 and five CpGs in SEU2) in order to assess the linearity of quantifying methylation ratios. As shown in Figure 5-20, both assays resulted in linear quantification (SEU1 - R^2=0.97 and SEU2 - R^2=0.99) indicating that there was no amplification bias observed. The average standard deviation obtained by duplicate analysis was 0.05, which relates to previous findings regarding other assays (5% for EFS assay in section 5.3.2.3). Methylation quantification was less accurate for partially methylated standards; however, since both loci are found either methylated or unmethylated in the analysed tissues, it was concluded that no methylation correction was needed. Obtained differences between observed and expected methylation could be explained by 'expected' amplification variations and pipetting errors in bisulphite PCR.

5.3.3.3.2 Sensitivity

The sensitivity of the proposed semen-specific methylation assays was assessed by analysing decreasing amounts of starting DNA material from blood (10 ng, 1 ng, 500 pg, 100 pg and 50 pg and 10 pg). Since initial development and validation of bisulphite Pyrosequencing® revealed the protocol's potential sensitivity as described in section 5.3.1.2., amounts down to 10 pg of DNA were bisulphite-treated in duplicate. The reason why non-semen DNA (blood) was analysed was to assess if false positive results would be obtained due to the low amounts of DNA used. As shown in Figure 5-21a, successful amplification and the expected blood methylation pattern for SEU2 was obtained down to 50 pg of starting DNA, which further corresponds to less than 10 cells given that each cell contains around 6 pg of DNA. This is very promising for the analysis of low-quantity or degraded samples, indicating the assay's applicability in

forensic casework. Bisulphite conversion did not seem to be affected by the starting DNA amount since an average of 91.5% conversion using all three controls included in the SEU2 assay was obtained for all amounts [Figure 5-21b]. It should be noted that for the first bisulphite control, conversion rates were higher (average of 96.2%) while the third control resulted in a mean rate of 85.4%. It is believed that this is due to low signal-to-noise ratio (low peak heights) or due to signal carry over from previous unincorporated cytosines. Similar results were obtained also for the SEU1 assay (data not shown).

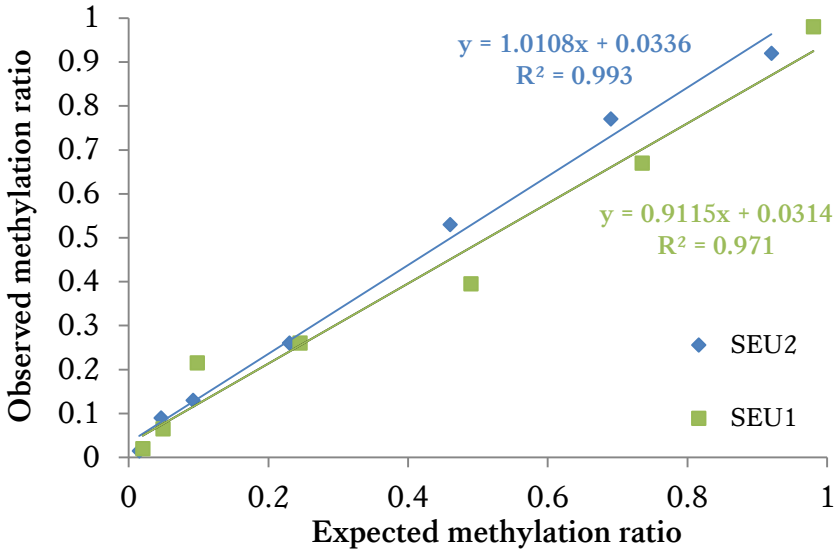


Figure 5-20. Linearity of methylation quantification for SEU1 and SEU2 assays
Standard curves showing the observed *vs.* expected average methylation ratio of each DNA methylation standard for both assays. As shown, for both assays methylation quantification was linear.

5.3.3.3.3 Aged samples

To further test the SEU1 and SEU2 assays' applicability in forensic casework, a set of aged and potentially degraded samples were tested. DNA from a set of nine semen samples was extracted shortly after collection and following storage at -20 °C for a year. Although the obtained DNA amount per extracted μl of semen significantly decreased [Figure 5-22], treating 10 ng of DNA the methylation status for all samples in both loci was found to be <0.2 methylated. No significant difference was obtained between the methylation values of fresh and stored semen samples ($p>0.05$). Lastly, a set of four semen stains on fabric (cotton) stored at -20 °C for 16 years were analysed using the proposed assays. Once again, no false negative results were obtained and the methylation patterns matched those obtained by freshly collected semen samples [Figure 5-19].

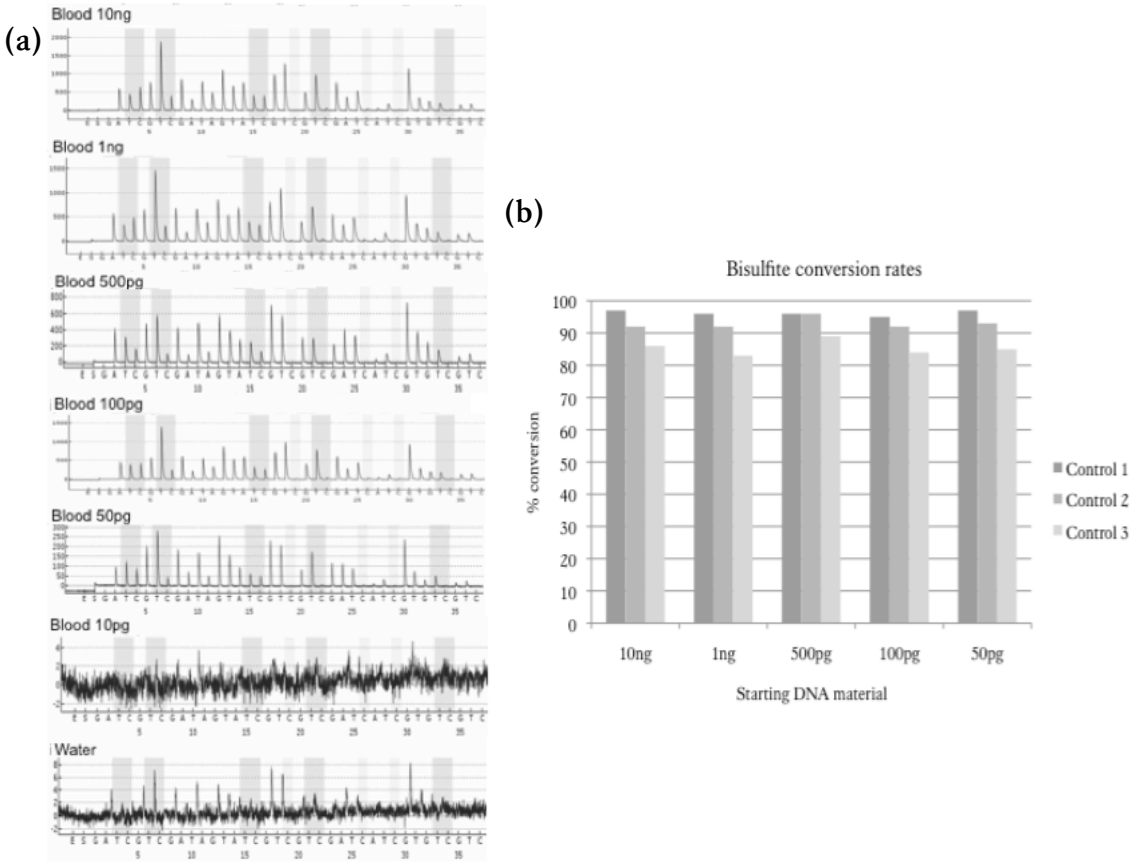


Figure 5-19. Sensitivity of SEU2 assay using decreasing amount of starting DNA material

(a) Obtained pyrograms™ when analysing SEU2 sequence using various starting blood DNA material (10 ng, 1 ng, 500 pg, 100 pg and 50 pg and 10 pg); as shown, successful amplification and the expected methylation pattern was obtained down to 50 pg, (b) observed bisulphite conversion rates for all three bisulphite conversion controls included in the assay.

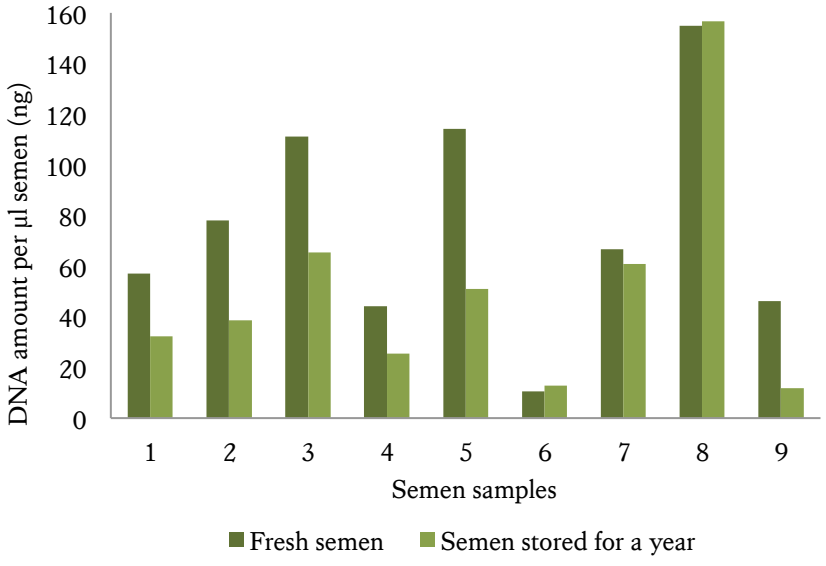


Figure 5-21. DNA recovery of the same semen samples shortly after collection and following a year of storage at -20 °C

5.3.4 Validation of immune cell-specific methylation markers

5.3.4.1 Optimisation of Pyrosequencing®-based assays

The third approach involved the investigation of six loci specific to cell types involved in the immune response including T cells, neutrophils, NK cells and naive CD8+, CD8A+ and CD8B+ T cells, previously reported to be differentially methylated among forensically relevant tissues [Figure 5-3]. For each assay, the corresponding DNA sequences were obtained from the EuroForGen collaborators highlighting the CpG sites included in their qPCR-based assays. Bisulphite Pyrosequencing® assays were designed using the BiSearch software and following the guidelines mentioned in section 2.2.4.1. However, it was necessary to optimise the PCR reactions to avoid mis-priming, primer self-annealing or the formation of non-specific PCR products. Each assay was optimised using an annealing temperature gradient, various concentrations of $MgCl_2$ and primer, as well as different PCR cycling conditions. Figure 5-23 shows a critical step of the optimization process where a set of ten annealing temperatures (50-63 ^{o}C) were tested using a blood sample.

5.3.4.2 Verification of methylation patterns

As illustrated in Figure 5-3, there were three assays showing semen-specific methylation (AMP1730, AMP2004 and AMP2007), two assays showing blood specific methylation (AMP1404 and AMP2007), while AMP1746 and AMP1817 demonstrated vaginal-specific methylation and AMP2004 could be used for the identification of menstrual blood. Rather than using individual assays per tissue, it was proposed that these six assays could be used together for the identification of the above-mentioned body fluids; however, it was believed that small methylation differences could also be used for the differentiation of saliva and skin (data not shown). To verify results obtained by qPCR-based assays (EuroForGen), two samples per tissue were analysed for all assays. Urine samples were also included in analysis to evaluate if they give false positive results for semen. The obtained results are shown in Table 5-7, where methylation ratios are colour-coded to help identifying variations among tissues. As illustrated, the results regarding the identification of semen were indeed confirmed with semen being completely methylated for AMP1730, unmethylated for AMP2007 and partially methylated in AMP2004.

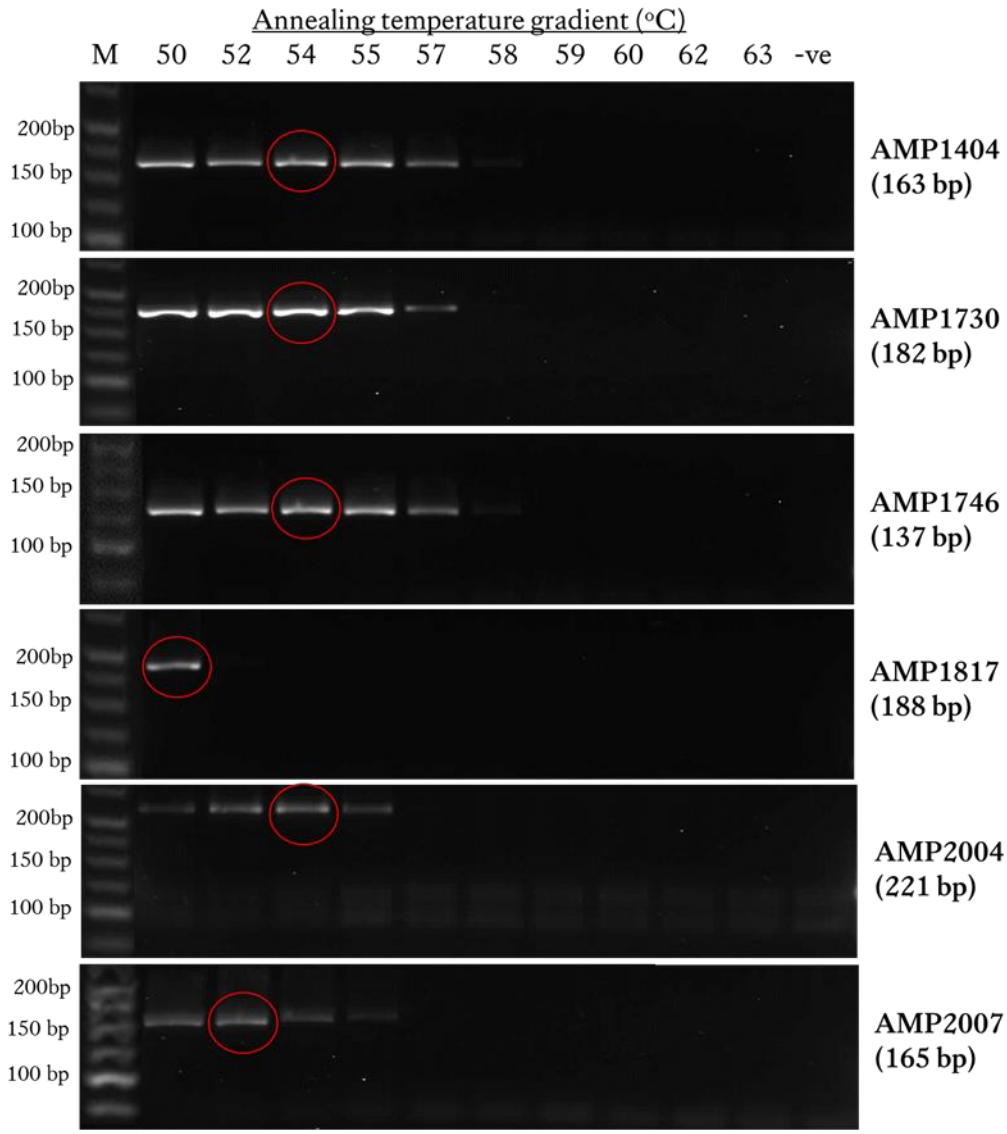


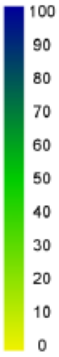
Figure 5-22. Final optimisation for all six immune cell-specific methylation assays
Agarose gel images showing the results of the final optimisation step (expected length in brackets) regarding annealing temperature (T_m). The first column represents the DNA marker (M) while the next ten columns show the resulted amplification bands of a blood sample for temperatures 50-63 °C . The last column represents the PCR negative (no-template) control. Red circles indicate the selected annealing temperature per assay that was believed to have shown better amplification efficiency.

Regarding blood, AMP1404 seemed to show blood-specific methylation pattern since blood was the only tissue demonstrating partial methylation. However, AMP2007 failed to give any differential profile for blood. Similarly, AMP1746 and AMP1817 seemed to be the least useful assays as all tissues exhibited partial or high methylation values [Table 5-7]. However, it was thought that CpG 6 in AMP1746 should be included in further analysis since both vaginal and menstrual secretion showed ~0.1 less methylation compared to the other tissues. Also, a skin sample showed lower methylation in CpG 4 in AMP1817 assay; therefore, further analysis would reveal if this was consistent among other skin samples.

Table 5-7. Verification of immune cell-specific methylation patterns by analysing two samples per tissue in each bisulphite Pyrosequencing® assay

Methylation values are colour-coded with blue indicating complete methylation, while yellow corresponds to unmethylated CpG sites. BL-blood, SA-saliva, SE-semen, VA-vaginal fluid, ME-menstrual blood, SK-skin, UR-urine.

Samples		BL		SA		SE		VA		ME		SK		UR	
		1	2	1	2	1	2	1	2	1	2	1	2	1	2
AMP1404	CpG1	0.76	0.65	0.99	1.00	1.00	1.00	1.00	0.98	0.98	0.95	1.00	0.88	1.00	0.86
	CpG2	0.69	0.58	0.95	0.92	0.93	0.90	0.89	0.92	0.87	0.90	0.65	0.90	0.93	0.84
	CpG3	0.66	0.58	0.99	0.97	0.96	0.96	0.86	0.90	0.87	0.90	0.97	0.96	0.89	0.86
	CpG4	0.46	0.44	0.60	0.50	0.57	0.61	0.67	0.59	0.61	0.60	0.67	0.51	0.50	0.58
	CpG5	0.58	0.53	0.84	0.78	0.85	0.84	0.86	0.80	0.83	0.83	0.32	0.69	0.75	0.69
AMP1730	CpG1	0.54	0.63	0.07	0.08	1.00	0.95	0.02	0.01	0.06	0.10	0.00	0.00	0.05	0.10
	CpG2	0.42	0.49	0.04	0.00	0.91	0.89	0.00	0.03	0.08	0.10	0.00	0.05	0.05	0.06
	CpG3	0.55	0.66	0.06	0.07	1.00	0.98	0.04	0.04	0.04	0.13	0.02	0.03	0.08	0.07
AMP1746	CpG1	0.86	0.91	0.99	0.97	0.99	0.97	0.94	0.95	0.89	0.90	0.99	0.95	0.97	0.96
	CpG2	0.91	0.92	0.95	1.00	1.00	1.00	0.96	0.96	0.94	0.92	0.89	0.92	0.94	1.00
	CpG3	0.88	0.91	0.93	0.94	0.94	0.93	0.94	0.92	0.91	0.92	0.96	0.96	0.89	0.94
	CpG4	0.75	0.81	0.77	0.85	0.86	0.79	0.77	0.75	0.70	0.72	0.93	0.82	0.72	0.83
	CpG5	0.73	0.77	0.74	0.74	0.76	0.78	0.74	0.71	0.71	0.66	0.80	0.77	0.71	0.76
	CpG6	0.83	0.85	0.86	0.86	0.85	0.90	0.80	0.76	0.77	0.74	0.93	0.85	0.81	0.84
	CpG7	0.75	0.77	0.82	0.80	0.77	0.80	0.81	0.79	0.80	0.74	0.85	0.75	0.78	0.79
AMP1817	CpG1	0.61	0.53	0.62	0.58	0.63	0.59	0.61	0.61	0.55	0.59	0.56	0.59	0.53	0.54
	CpG2	0.58	0.48	0.56	0.50	0.52	0.55	0.55	0.56	0.48	0.50	0.41	0.59	0.50	0.49
	CpG3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	CpG4	0.88	0.84	0.85	0.90	0.90	0.95	0.89	0.93	0.90	0.87	0.92	0.76	0.92	0.88
	CpG5	0.39	0.34	0.47	0.34	0.41	0.41	0.40	0.39	0.34	0.39	0.37	0.45	0.35	0.39
	CpG6	0.89	1.00	0.96	0.95	0.94	1.00	1.00	0.97	1.00	0.95	1.00	1.00	1.00	1.00
	CpG7	0.94	0.87	0.84	0.83	0.87	0.91	0.85	0.86	0.77	0.83	0.61	0.86	0.87	0.77
	CpG8	0.37	0.41	0.41	0.46	0.55	0.42	0.47	0.45	0.34	0.45	0.55	0.58	0.60	0.52
	CpG9	0.92	0.85	0.85	0.88	0.76	0.90	0.84	0.86	0.77	0.85	0.77	0.76	0.86	0.78
	CpG10	0.54	0.45	0.53	0.53	0.64	0.43	0.56	0.47	0.30	0.46	0.28	0.41	0.96	0.29
	CpG11	0.79	0.81	0.75	0.76	0.76	0.82	0.71	0.69	0.60	0.64	0.57	0.77	0.70	0.64
AMP2004	CpG1	0.49	0.39	0.64	0.66	0.83	0.75	0.81	0.77	0.79	0.82	0.33	0.74	0.72	0.64
	CpG2	0.84	0.80	0.99	0.99	0.75	0.75	0.98	0.96	1.00	0.97	0.98	1.00	0.98	0.99
	CpG3	0.86	0.78	1.00	0.98	0.60	0.58	1.00	1.00	0.96	1.00	1.00	1.00	0.99	0.99
	CpG4	0.29	0.27	0.37	0.35	0.18	0.19	0.30	0.31	0.24	0.28	0.46	0.38	0.38	0.35
	CpG5	0.76	0.67	0.88	0.85	0.71	0.66	0.88	0.84	0.74	0.87	0.57	0.88	0.84	0.75
AMP2007	CpG1	0.87	0.85	0.89	0.84	0.04	0.16	0.63	0.51	0.66	0.73	0.87	0.95	0.81	0.81
	CpG2	0.51	0.28	0.26	0.23	0.01	0.09	0.47	0.16	0.48	0.36	0.64	0.28	0.86	0.33
	CpG3	0.37	0.11	0.08	0.11	0.01	0.07	0.38	0.08	0.38	0.16	0.34	0.09	0.69	0.09
	CpG4	0.84	0.86	0.93	0.94	0.12	0.15	0.86	0.83	0.86	0.92	0.97	0.92	0.94	0.90
	CpG5	0.84	0.83	0.82	0.84	0.12	0.16	0.74	0.71	0.75	0.79	0.77	0.75	0.81	0.79
	CpG6	0.83	0.83	0.84	0.88	0.17	0.14	0.77	0.73	0.78	0.85	0.92	0.81	0.88	0.81
	CpG7	0.76	0.78	0.61	0.63	0.17	0.13	0.72	0.29	0.70	0.73	0.83	0.46	0.81	0.61
	CpG8	0.84	0.83	0.89	0.92	0.25	0.16	0.82	0.79	0.80	0.88	0.86	0.90	0.86	0.88



5.3.4.3 Specificity of selected markers

To further test the specificity of the proposed assays and to reduce the analysis cost since in most assays two separate sequencing reactions were required to analyse all CpG sites [Table 5-5], the most informative CpG sites per assay that could be combined in one sequencing reaction were selected for further testing. These included CpGs 1-3 for AMP1404, CpGs 1-2 for AMP1730, CpGs 1-6 for AMP1746, CpGs 1-5 for AMP1817, CpGs 2-3 for AMP2004 and CpGs 4-6 for AMP2007.

To validate the specificity of the selected CpG sites and since for some of them the methylation differences were quite small (~10%), it was decided that a larger dataset should be analysed including around 20 samples per tissue. Also, since one the EuroForGen researchers involved in this collaborative project reported the growing need to identify nasal blood (mainly in cases of physical assault), six DNA samples from nasal blood as well as twelve nasal fluid samples were also co-analysed. Although it is appreciated that identifying nasal fluid would be rare in a forensic scenario, it was analysed for comparison with the other tissues (nasal blood/epithelial tissues) and assess if there are any differences. Therefore, a total of 144-154 tissue samples of both females and males of different ages were analysed for each methylation assay. Since qPCR assays were based on the co-analysis of all CpG sites included in each assay and also since similar methylation levels were obtained among each assay's CpGs in the verification experiment, potential tissue-to-tissue variations were studied utilising the average methylation per locus [Figures 5-22 and 5-23]. Methylation differences using individual CpG sites were also examined (e.g. CpG 4 in AMP1817) but did not seem to be any different from when using the mean ($p>0.05$).

Firstly, in the proposed blood-specific AMP1407 locus, whole blood demonstrated methylation levels ranged from 0.46-0.90 (mean=0.7), while most of the other tissues exhibited higher methylation levels with the exception of skin and urine, where significant variation in methylation levels was observed. One can argue that differentiating between blood and urine or skin can be supported by visual examination of the stain, caution is needed when conclusions regarding the tissue type present are made using this assay.

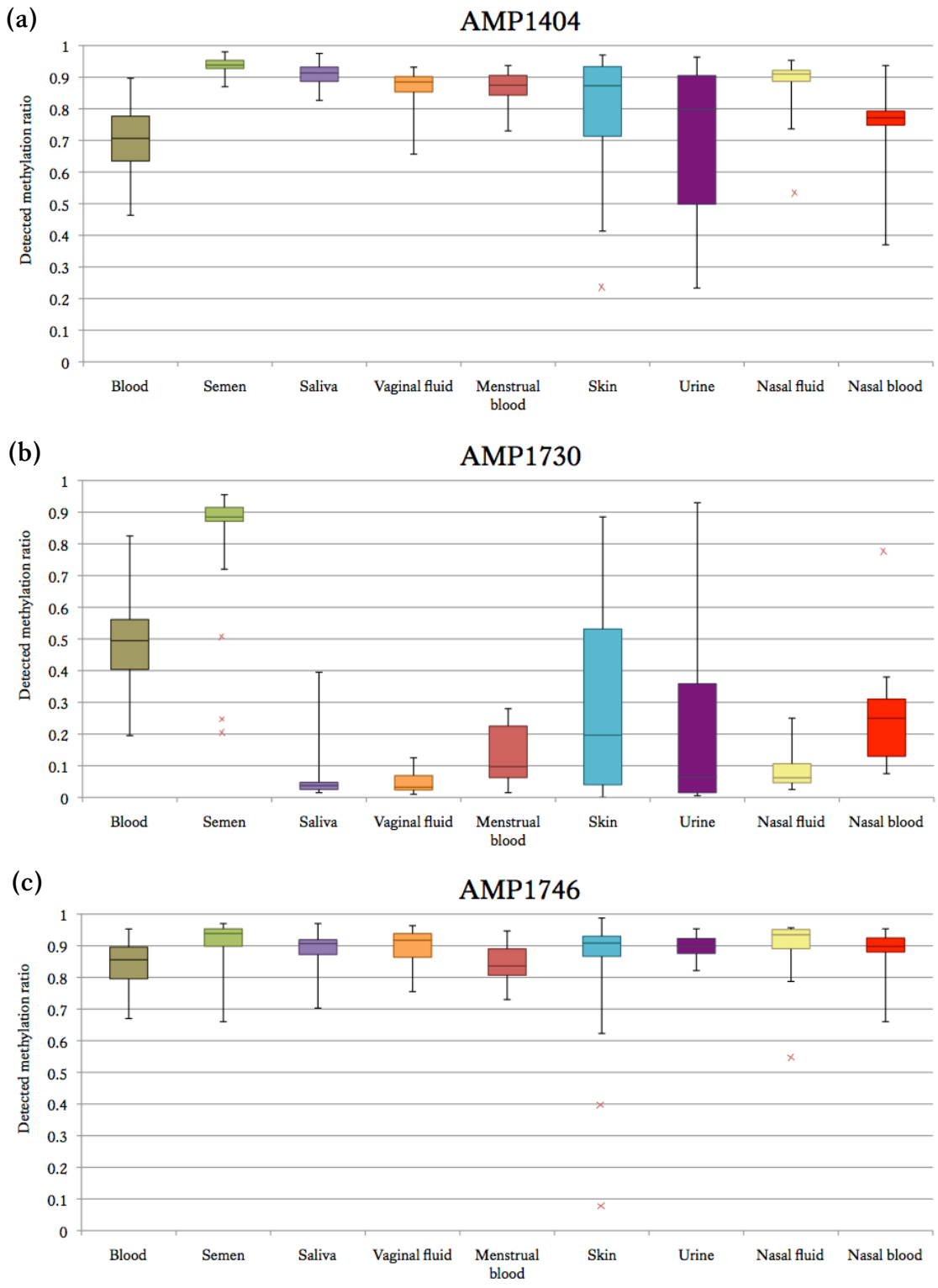


Figure 5-23. Methylation levels of (a) AMP1407, (b) AMP1730 and (c) AMP1746 in various forensically relevant body fluids/tissues

Observed mean methylation ratio in various body fluids/tissues. Data are presented in the form of box-and-whisker plots showing the first and third quartiles (boxes), the median (horizontal line) and minimum and maximum (error bars) methylation values detected. Outliers (methylation value ≥3*SD) are shown separately as red 'x' dots.

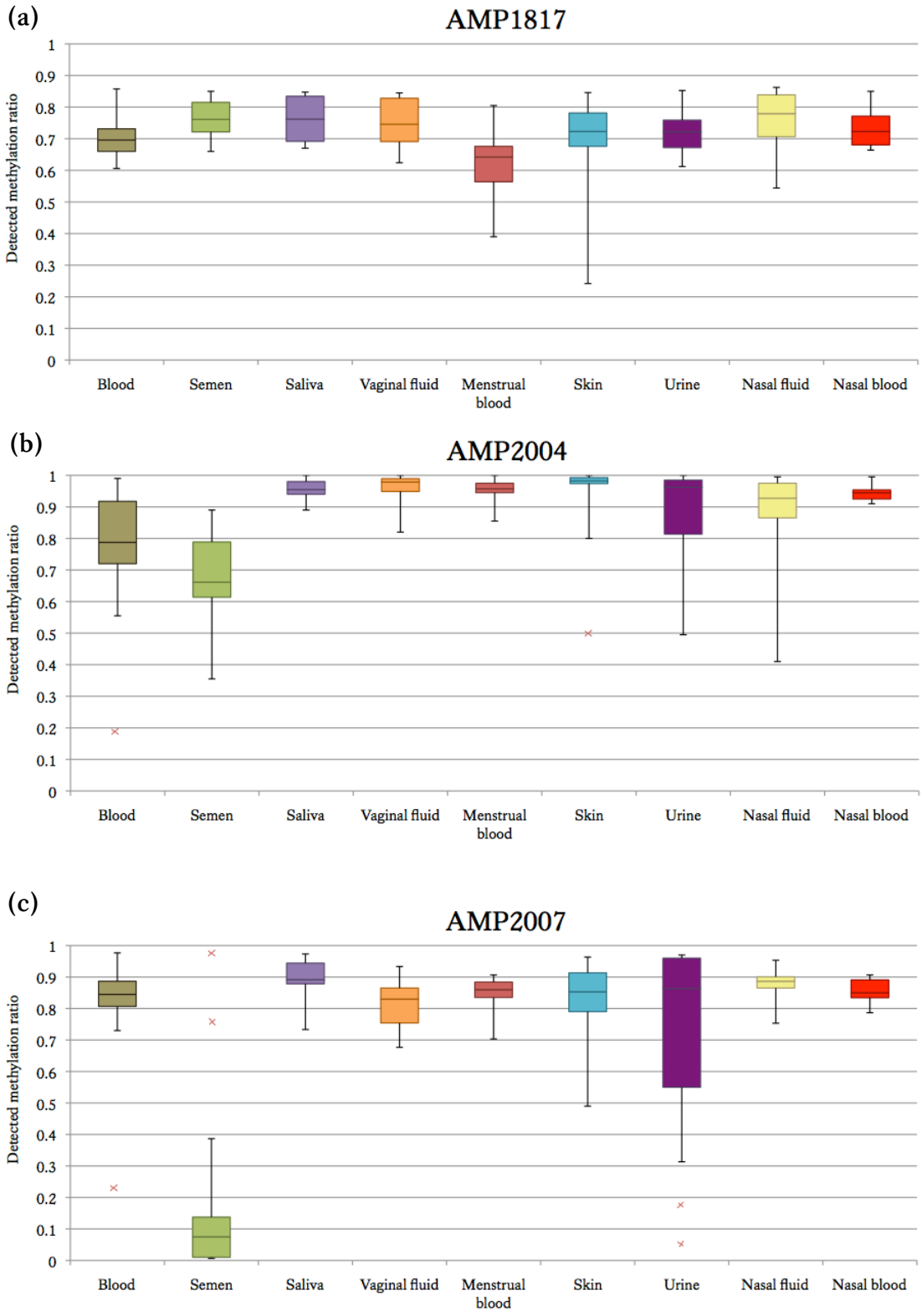


Figure 5-24. Methylation levels of (a) AMP1817, (b) AMP2004 and (c) AMP2007 in various forensically relevant body fluids/tissues

Observed mean methylation ratio in various body fluids/tissues. Data are presented in the form of box-and-whisker plots showing the first and third quartiles (boxes), the median (horizontal line) and minimum and maximum (error bars) methylation values detected. Outliers (methylation value ≥3*SD) are shown separately as red 'x' dots.

With regards to AMP1730 and AMP2007 loci, we can conclude that semen-specific DNA methylation profiles were obtained in both assays; however, there were two semen in particular demonstrating the 'opposite' methylation profile (out of 20). These two samples are the same ones that have previously been highlighted due to their low sperm count indicating potential male infertility issues [Figure 5-17 & 5-19]. Since these two samples have consistently given the 'opposite' methylation profile, the need to include the possibility of disease status while interpreting methylation profiles is further highlighted. Interestingly, for both assays the urine samples gave positive semen profile; however, it is thought that this could be due to 'contamination' of urine with semen or epithelial cells (since skin samples resulted in similar profiles with urine). Most of these urine samples were obtained from male volunteers (86% for AMP1730 and 57% for AMP2007), a fact that further supports the above statement.

Additionally, although AMP2004 seemed to be useful for differentiation of semen and blood (partially methylated) from other body fluids (highly methylated) in the verification experiment, analysing more samples revealed greater variations in methylation levels. It is concluded that this marker cannot be applied in forensic casework with confidence, since for example as shown in Figure 5.24b, a sample generating methylation of 0.8 could be any of the following: blood, semen, skin, urine or nasal fluid. Finally, the assays AMP1746 and AMP1817, although initially reported as vaginal-specific markers [Figure 5-3], they failed to reproduce this result in the present study. All tested body fluids showed similar methylation levels for both assays.

5.4 Final remarks

In this study, DNA methylation profiling for use in detecting body fluids seemed to be very promising. Initial evaluation of a bisulphite Pyrosequencing®-based assay (HBA1) revealed that this method can be highly sensitive (successful DNA methylation profiles down to 100 pg) and reproducible (average of 5% standard deviation). However, it was noticed that PCR amplification bias were present resulting in a 'curved' linearity graph, supporting the need to 'correct' detected methylation values. Subsequently, a set of 18 genomic loci were selected using three different approaches and tested among forensically relevant body fluids and tissues. In general, it was observed that semen and blood demonstrated differential DNA methylation patterns and was easy to detect with confidence; however, more complex tissues such as saliva, vaginal fluid and menstrual blood were more challenging using the proposed markers. Analysing more samples in the future could potentially reveal if observed variations in methylation are due to inter-individual changes or other factors. When testing mock casework samples, the tissue-specific markers seemed to be applicable and DNA methylation profiles seemed to be stable for up to 18 years in blood and 16 years in semen stains.

Part 2

6 Literature review on estimating the chronological age of an individual

The ability to accurately estimate a person's chronological age would be a great advantage in police investigations as it could provide significant investigative leads. In a forensic context, predicting age is necessary for both the dead and the living. Predicting the age of an unidentified cadaver (age at death) could assist in the personal identification process (for example, in mass disasters) by creating a biological profile that can potentially be compared to missing persons. For the living, age prediction could be used to solve judicial or civil issues concerning age of minors or adults that lack valid identification documents or are involved in cases of adoption. Moreover, for intelligence purposes, estimating the age of a crime scene stain's donor could potentially narrow down the number of suspects, especially in cases where an eye witness is not available.

There have been various approaches to estimate age at death of human remains or chronological age of living individuals that will be described in this chapter; however, a common problem is the lack of standardisation of methods and sampling. A fundamental assumption of most methods is that the biological age of a person corresponds to their chronological age. However, while the chronological age is the calendar age which is usually identified in years, the biological age refers to how ageing affects the body and how this might be recognised. Naturally, the older the person is the larger might be the discrepancy between these two; therefore, age estimation is usually less accurate for older individuals. Most of these attempts are based on alterations of tissues or organs at the molecular level as a result of the natural process of ageing; however, the developed methods are usually relative producing an estimate with a large age range.

Undoubtedly, developing an age prediction model is a major challenge for forensic scientists since they would need to be able to apply and validate it using minute or degraded samples consisting of a range of tissues and body fluids. This chapter will review previously proposed age prediction methods in literature and will discuss their advantages and drawbacks in terms of use in a forensic scenario.

6.1 Relevant background

Ageing is a very complex process influenced by various genetic, lifestyle and environmental factors. It causes a variety of modifications and adjustments in tissues and organs that accumulate over an individual's lifetime. In a medical setting, these age-related factors provide useful information for 'personalised' medicine as they are used as risk factors in the development of many diseases, such as cancer (Teschendorff *et al.*, 2010; Tsai & Baylin, 2011). In a forensic scenario, and in cases where a human skeleton is recovered, a rough estimation of the age-at-death can often be achieved through various morphological age-associated changes in the skeleton or in the dentition (Lynnerup *et al.*, 2010). While soft tissue may also be present, forensic anthropologists rely on odontological and skeletal maturity to estimate age (Cunha *et al.*, 2009). Nevertheless, difficulties arise when skeletons are incomplete and even though scientists have focused on studying particular parts of the body (e.g. the acetabulum) (Rouge-Maillart *et al.*, 2007), estimation is still not accurate.

Even though estimating age through skeletal structures is very common, efforts have been made to develop a potentially more accurate molecular-based test. Over the course of a lifetime, stochastic events lead to gradual changes of biomolecules that can be examined and characterised as both the cause and the consequence of our 'molecular clock'. There are various age-related mechanisms that have been thoroughly studied over the last two decades and could provide promise in their application within a forensic scenario. These include various chemical modifications, gene expression alterations and variations at the DNA level (Meissner & Ritz-Timme, 2010; Zapico & Ubelaker, 2013).

The chemical methods include protein modifications like the aspartic acid racemisation (Arany *et al.*, 2004) and formation of advanced glycation end-products (Pilin *et al.*, 2007), accumulative exposure to chemicals like lead (Al-Qattan & Elfawal, 2010) and also, alterations in cell components like collagen crosslinks (Martin-de las Heras *et al.*, 1999). On the other hand, the molecular biology methods include gene expression changes (Glass *et al.*, 2013), alterations in the composition of metabolomic markers (Menni *et al.*, 2013), accumulation of DNA damage as reflected through the amount of deletions seen in mitochondrial DNA (Meissner *et al.*, 1997),

shortening of telomeres that occurs along with cell division (Tsuji *et al.*, 2002) and lastly, decrease in sjTREC rearrangements (Ou *et al.*, 2011).

6.1.1 Chemical methods

6.1.1.1 Lead accumulation

Lead is one of the most common and important environmental pollutants and its concentration in blood can act as an indicator of immediate exposure (Al-Qattan & Elfawal, 2010). On the other hand, lead concentration in teeth, particularly in dentine, is thought to point out long-term exposure and has been used in the past as a measure of lead pollution (Steenhout & Pourtois, 1981). Generally, the correlation between lead concentration in teeth and age has not yet been well established. However, Al-Qattan and Elfawal managed to build an age prediction model using lead accumulation in teeth of the Kuwaiti population that could estimate age with a mean error of 1.3±4.8 years. Authors observed differences in prediction between males and females as males demonstrated a higher lead accumulation. Although these results seem promising, further research is required to apply similar formulas to other populations.

6.1.1.2 Collagen crosslinks

The collagenous matrices of skeletal connective tissues are fixed by covalent crosslinks between collagen molecules via intermolecular reactions of aldehyde resides (Eyre, 1987). There are mainly two crosslink pathways, the lysine aldehyde pathway in skin and the hydroxylysine aldehyde pathway in bone and cartilage. These crosslinks have been reported to fade as connective tissue matures by converting borohydride-reducible aldimines to mature non-reducible compounds such as pyridinoline. This maturation process is believed to be age-related since a decrease in the reducible crosslinks with increasing age has been observed in various tissues (Bailey & Shimokomaki, 1971). In a forensic approach, a component of non-reducible crosslinks, namely deoxypyridinoline (DPD) was studied in permanent molars from individuals aged 15 to 73 years old (Martin-de las Heras *et al.*, 1999). Using an enzyme immunoassay method, authors measured the DPD ratio with respect to age and obtained an estimated error of ±14.9 years at a 65% confidence level. Interestingly, deviations between different age groups were observed.

6.1.1.3 Aspartic acid racemisation (AAR)

In mammals, it is known that only L-amino acids are integrated during protein synthesis. However, the chemical stability of some residues, aspartyl and asparaginyl residues in particular, has been reported to decrease with age leading to post-translational non-enzymatic protein modifications (Lowenson & Clarke, 1988). Racemisation is a natural process, where active compounds are converted into a racemic mixture causing changes in the biological activities or chemical properties of proteins. These alterations can be measured as an age-dependent increase of the D-aspartic acid content in acid hydrolysates of the studied proteins, which is believed to be a result of aspartic acid racemisation. Although the mechanisms involved seem to be quite complex, the relationship between AAR and age is thought to be very close for permanent proteins. The optimal tissue for this type of study is once again the tooth dentine; however protocols have also been developed for bone and skin as in theory any tissue that contains metabolically-stable proteins can be used for age estimation (Ritz-Timme *et al.*, 2003).

Significant disadvantages of such a method include the need of several teeth of the same kind, the fact that racemisation might also occur post-mortem, especially if the human remains have been exposed to high temperatures and also the presence of bacteria. Despite these drawbacks, several studies have demonstrated the precision of the method (mean error of 1.5-4 years) (Dobberstein *et al.*, 2010; Ohtani & Yamamoto, 2010); however, one has to have in mind that the reliability of the results strongly rely on post-mortem environmental conditions (Dobberstein *et al.*, 2008).

6.1.1.4 Advanced glycation end-products (AGEs)

Similarly to racemisation, non-enzymatic reactions between carbohydrates and proteins known as Maillard reactions produce a number of age-related protein modifications. These reactions are related to browning, fluorescence and cross-linking of proteins. Advanced glycation end-products (AGEs) accumulate in long-lived proteins like tissue collagens and can promote age-related conditions such as diabetes (Thorpe & Baynes, 1996). The accumulation of these non-enzymatic glycation products is verified by the colour change of articular cartilage from blue in young age to yellow/brown in the elderly. The importance of colour changes was emphasized by

investigating age-related colour changes in intervertebral discs excisions, Achilles tendons and rib cartilage caused by the formation of AGEs (Pilin *et al.*, 2007). Furthermore, using spectroradiometry, Martin-de las Heras *et al* (2003) analysed different colorimetric variables in teeth obtaining an average error of 13.7 years. However, this technique is not suitable for specimens after extended post-mortem intervals since differences in dental colour were observed.

In conclusion, the proposed chemical methods mentioned in this section demonstrate either low age prediction accuracy or can be mainly applied in connective tissues and skeletal or dental remains. Therefore, these methods are primarily for use in cases where age prediction is required in post-mortem specimen and should be applied with caution.

6.1.2 Molecular biology methods

6.1.2.1 Metabolomic markers

Metabolomics is a novel technology which aims to profile all low-molecular-weight metabolites present in a biological sample that could be correlated with various physiological and pathophysiological processes (Psychogios *et al.*, 2011). It is therefore a useful platform for investigating in a single approach all possible ways that metabolism is affected by a certain variable such as age. Changes in protein, energy and lipid metabolism have been observed with increasing age while these metabolites are also statistically linked with sex and race (Lawton *et al.*, 2008). In a much wider approach, a study using 2,162 healthy individuals (32-81 years old) investigated a large set of metabolites via Flow Injection Analysis/Mass Spectrometry (FIA-MS/MS) and showed that metabolic profiles are strongly correlated with age (Yu *et al.*, 2012).

A total of 71 metabolite concentrations in women and 34 metabolites in men were significantly associated with age; five of them were common in both genders. All reported metabolites are involved in major biological functions such as altered cellular membrane composition, mitochondrial metabolism and oxidative stress. However, sexual dimorphisms were extensively observed and authors suggested that metabolomic variations should be analysed separately for women and men. Also, no

conclusions could be made whether these changes were a result of ageing, other physiological aspects or as part of responses to damaging agents (Yu *et al.*, 2012).

Similarly, 280 different metabolites were investigated in a data set of 6,055 individuals (17-85 years old) and a panel of 22 age-associated metabolites was identified (Menni *et al.*, 2013). These metabolites included nine lipids, seven amino acids, two intermediates in the energy pathway, two xenobiotics, one carbohydrate and one nucleotide. One metabolite in particular, namely C-glycosyl tryptophan (C-glyTrp) correlated strongly not only with age, but also with lung function, bone mineral density and weight at birth. Using a Cox regression model, researchers also performed a test for association with mortality as a function of the linear combination of all 22 metabolites [Figure 6-1]. The derived variable was used as the independent variable, while potential death was the outcome variable (mean follow-up time 7.33±4.46 years).

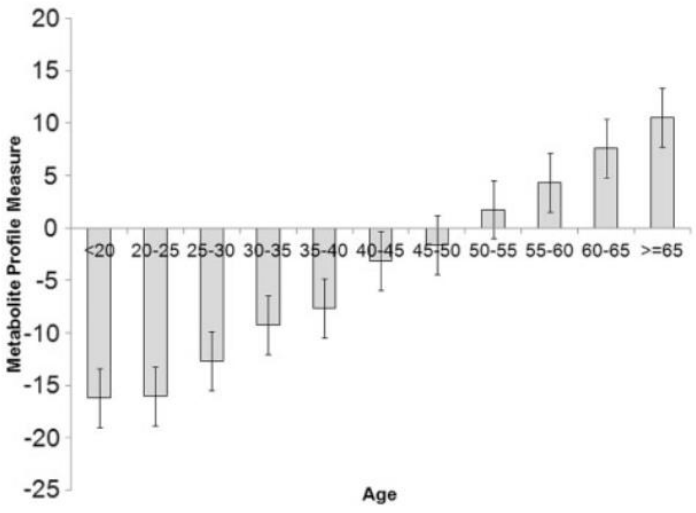


Figure 6-1. Metabolic profile measures and age (Menni *et al.*, 2013)
Using the coefficients from the stepwise regression on age of the 22 metabolites, a metabolic profile measure was obtained for each individual taking part in the study. The mean and standard error of this variable was computed for intervals of five years of age.

6.1.2.2 Gene expression patterns

Several gene expression studies of ageing and ageing-associated disorders have revealed that there are various genes that are turned on or off in an age-dependent manner. For example, a study in post-mortem human brain from 30 individuals aged 26-106 years old revealed that almost 4% of the 11,000 analysed genes demonstrated age-related expression patterns in individuals aged above 40 years (Lu *et al.*, 2004).

Likewise in kidney, 985 age-associated genes were identified after analysing 74 individuals aged 27-92 years old (Rodwell *et al.*, 2004). In an attempt to determine genes and pathways involved in age in multiple tissues, the gene expression of skin, adipose tissue and various lymphoblastoid cell lines (LCLs) from 856 female twins aged from 39-85 years old was investigated by Glass and his colleagues (2013). Most age-associated genes were linked to fatty acid metabolism, mitochondrial activity, cancer and splicing; however, significant tissue-specific differences were evident (Glass *et al.*, 2013). Interestingly, LCLs lacked age-related genes suggesting that the transformation to an immortalised cell masks (or even removes) any age-related gene expression signatures.

Moreover, from a forensic point of view the analysis of tissue cytochrome c oxidase (CCO) activity as well as the quantification of its protein content and mRNA expression was proposed (Ishikawa *et al.*, 2011). As expected, CCO activity was found to gradually decrease with age in the human heart tissue (r=0.83). Nevertheless, one has to take into account that the measurement of CCO activity is only useful in cases where samples can be collected shortly after death.

6.1.2.3 Mitochondrial DNA (mtDNA) deletions

It is known that mitochondria are semi-autonomous organelles having their own protein synthesis and DNA replication machineries. Mitochondrial processes mainly involve five enzyme complex mechanisms essential for respiration and oxidative phosphorylation, where lipids and glucose are oxidised to produce ATP. During this procedure 0.2% of the utilised oxygen is released as free radicals like hydrogen peroxide or superoxide, which are highly reactive oxygen species causing damage to proteins, lipids and especially mitochondrial DNA (Mandavilli *et al.*, 2002). Mitochondrial mutagenesis is mostly found in the form of stochastic deletions of mtDNA that show a tissue-specific accumulation pattern and usually increase with advancing age (Meissner *et al.*, 2006).

The 4,977 bp deletion, also known as the common deletion, has been analysed for forensic applications in various tissues using PCR-based assays. A study on skeletal muscle revealed a correlation (r=0.83) between the frequency of the 4,977 bp deletion and the age at death (Meissner *et al.*, 1999). However, scientists were unable to

reproduce the same results in whole blood (Mohamed *et al.*, 2004). Some of the main challenges of applying such method for age estimation include the exponential nature of PCR that could lead to differences within the same sample, the significant differences that have been observed among tissues as well as possible mosaicism even between neighbouring cells (Storm *et al.*, 2002). Further research is required to assess if the 4,977 bp could serve as an alternative for age estimation since external factors such as pathologic conditions (hypoxia, ischemia and sudden cardiac death) have been reported to affect the rate of deletion generation, more likely due to increased production of free radicals (Polisecki *et al.*, 2004).

6.1.2.4 Telomeres

Telomeres have been recognised for their role in cell survival and replicative capability of dividing somatic cells; without telomeres cells would not be able to differentiate between DNA breaks within the genome and the chromosome ends (Jiang *et al.*, 2007). Telomeres are found at the end of linear chromosomes consisting of thousands of tandem repeat units and are stabilised by telomere-binding proteins. This telomere capping is responsible for maintaining chromosomal stability and preventing cell cycle arrest (de Lange, 2004). However, due to the end-replication problem of DNA polymerase, or processing of telomeres during the cell cycle, telomere shortening is common after each cell division and can only be fixed by inducing the expression of telomerase (von Figura *et al.*, 2009).

The use of telomere length for biological age prediction has been explored in blood, where it has been observed that the mean terminal restriction fragment length decreases 20-60 bp every year during a human life (von Zglinicki & Martin-Ruiz, 2005). In another study, using Southern blot analysis the correlation between telomere length and age was significant (r=0.83); however, there were cases where young and old individuals shared the same length (Tsuji *et al.*, 2002). Therefore, telomere length could more likely be used to assign an age interval rather than a specific age demonstrating limited accuracy. These variations are suspected to be a result of genetic and environmental factors such as oxidative stress (von Figura *et al.*, 2009; Von Zglinicki, 2000). Differences in telomere attrition rate between different chromosomes are also evident (Britt-Compton *et al.*, 2006).

6.1.2.5 sjTREC rearrangements

Another mechanism that has been proposed to reliably estimate biological age is the use of T-cell DNA rearrangements (Ou *et al.*, 2011; Zubakov *et al.*, 2010). T lymphocytes possess specific receptors encoded by the T-cell receptors (*TCRs*) genes to recognise unknown antigens. To create a wide range of TCR molecules, each T cell experiences distinctive somatic rearrangements at these particular loci, where intervening DNA sequences are deleted forming episomal DNA molecules called signal joint TCR excision circles (sjTRECs) (Breit *et al.*, 1997). It is believed that the number of sjTRECs declines as age increases.

Zubakov *et al* (2010) developed a robust and sensitive (down to 5 ng of starting DNA material) real-time quantitative PCR protocol using 195 whole blood samples derived from healthy Dutch individuals ranging in age from a few weeks to 80 years. However, the standard error of the estimate remained high at ±8.9 years [Figure 6-2]. Storage time of blood samples did not seem to affect age prediction, however statistically significant gender differences on sjTREC quantification were observed. Applying a similar approach, another study was carried out in 248 Chinese Han individuals aged between 0 (new-borns, cord blood) to 78 years old (Ou *et al.*, 2011). The authors were able to obtain a correlation coefficient of r=-0.82 by regression analysis between sjTREC levels and age with 65.3% of age estimates lying within 10 years of the actual age of the subjects. It is very likely that conditions affecting the immune system's health status such as HIV/AIDS or leukaemia could further affect the proposed age estimation. It should also be noted that such a test is restricted to blood and it would not be possible to apply this method in other tissues such as semen or saliva since they do not contain T cells in quantities that would allow for sjTREC detection.

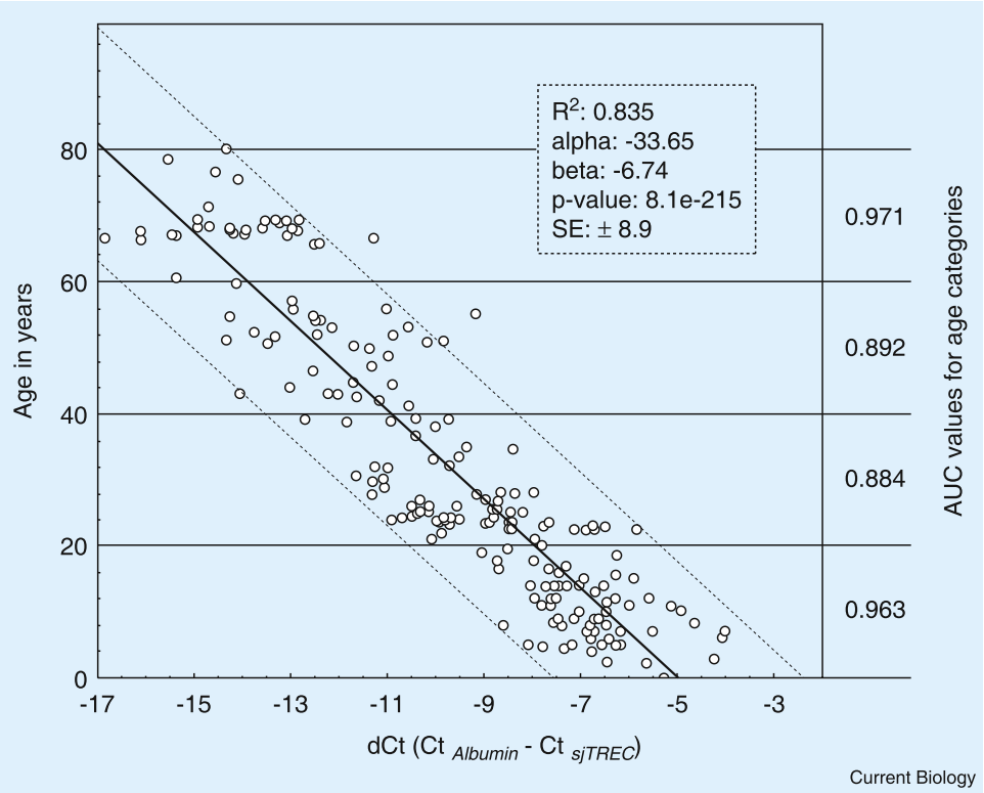


Figure 6-2. Age prediction in blood using sjTREC abundance (Zubakov *et al.*, 2010)
Linear regression model of the relationship between chronological age and normalised sjTREC numbers in whole blood. Dotted lines correspond to 95% prediction interval.

All methods mentioned above show limitations. Tests based either on chemical or molecular biology protocols so far are more likely to suggest an age group (generation) rather than accurately predict age. Some of these tests could result in the destruction of the original sample, limiting downstream analysis and therefore their applicability in crime scene samples. While some of these tests, especially methods based on chemical modifications, can be applied in predicting age using human remains, in this study predicting age by examining bodily fluids of living individuals was desired.

6.2 Age-associated DNA methylation

Epigenetic analysis could serve as an alternative or supplementary method for age prediction since particularly DNA methylation is well-known to be one of the mechanisms responsible for cell differentiation and the cellular response to ageing (Bell *et al.*, 2012; Day *et al.*, 2013). It is generally suggested that there is an increase in global epigenetic drift with age (Teschendorff *et al.*, 2013) and various genome-wide methylation analyses have revealed a substantial decrease in global DNA methylation levels with advancing age (Gentilini *et al.*, 2013). Comparing new-born and centenarian genomes it could be seen that the centenarian DNA had a lower DNA methylation content with a reduced correlation in the methylation status of neighbouring CpG sites across the genome in comparison with the more homogenously methylated new-born DNA (Heyn *et al.*, 2012). These CpGs covered all genomic compartments including promoters, exonic, intronic and intergenic regions.

Changes in DNA methylation patterns due to ageing are quickly observed during the first months of an individual's life and throughout childhood (Alisch *et al.*, 2012; Martino *et al.*, 2013). Cumulative evidence points towards the distinct contributions of genetic (Bell *et al.*, 2011), environmental (Gronniger *et al.*, 2010; Lee & Pausova, 2013) and stochastic factors to DNA methylation levels at single genomic areas. However, as shown in previous chapters, epigenetic changes can also be tissue-specific (Thompson *et al.*, 2013). Genome-wide methylation changes have been investigated in various forensically relevant tissues including blood (Johansson *et al.*, 2013), skin (Koch *et al.*, 2011; Raddatz *et al.*, 2013), brain (Horvath *et al.*, 2012) and skeletal muscle (Zykovich *et al.*, 2014).

6.2.1 Age-associated CpG sites in blood

In order to identify specific age-associated differentially methylated CpG sites for a particular body fluid, Horvath *et al* (2012) have chosen to perform genome-wide studies that enable analysis of thousands of CpGs at the same time. It is believed that monozygotic twins serve as an ideal model to study these dynamic epigenetic marks as they share the same DNA sequence and start life with almost identical methylation patterns (Fraga *et al.*, 2005; Li *et al.*, 2013). In a recent study, the heritability and

relationship with age and gender of selected DNA methylation profiles was investigated using genomic DNA derived from whole blood of various twin pairs (Boks *et al.*, 2009). Genes that contained age-associated CpG sites were identified including activin A receptor type I (*ACVR1*), interleukin 6 (*IL6*), caspase recruitment domain-containing protein 15 (*CARD15*), platelet-derived growth factor receptor alpha (*PDGFRA*), nuclear factor kappa-B subunit 1 (*NFKB1*) and the ETS-domain protein (*ELK*) gene. Most of these genes were also included in the list of the top 100 age-associated CpG sites identified by Christensen and his team, where they also showed that these alterations are dependent upon genes' CpG island context (Christensen *et al.*, 2009).

Utilising Illumina's 27K array which allows determination of bisulphite conversion-based, single-CpG resolution, DNA methylation levels at 27,578 different CpG sites within more than 14,000 promoters in the human genome, Rakyan *et al* (2010) observed a total of 213 CpG sites that become more methylated with age and another 147 CpGs that lose methylation over time in blood. Interestingly, >95% of these sites were located within 500 bp of the transcriptional start site of the associated gene, implying a connection with regulation of gene expression. In an attempt to correlate blood age-associated DNA methylation changes with other tissues, Horvath and his colleagues analysed 2,442 Illumina DNA methylation arrays from brain and blood tissues. Consensus module analysis revealed common and robust co-methylation relationships suggesting that age effects on DNA methylation levels are well-preserved between these two tissues (Horvath *et al.*, 2012).

Additionally, another three promising epigenetic markers of age were identified by analysing 494 whole blood samples from individuals aged 9-99 years old using Illumina's 450K array (Garagnani *et al.*, 2012). These included the ELOVL fatty acid elongase 2 (*ELOVL2*), four and a half LIM domains 2 (*FHL2*) and perproenkephalin (*PENK*) genes. For all three sites, methylation levels increased with age with ELOVL2 displaying the widest methylation range [Figure 6-3].

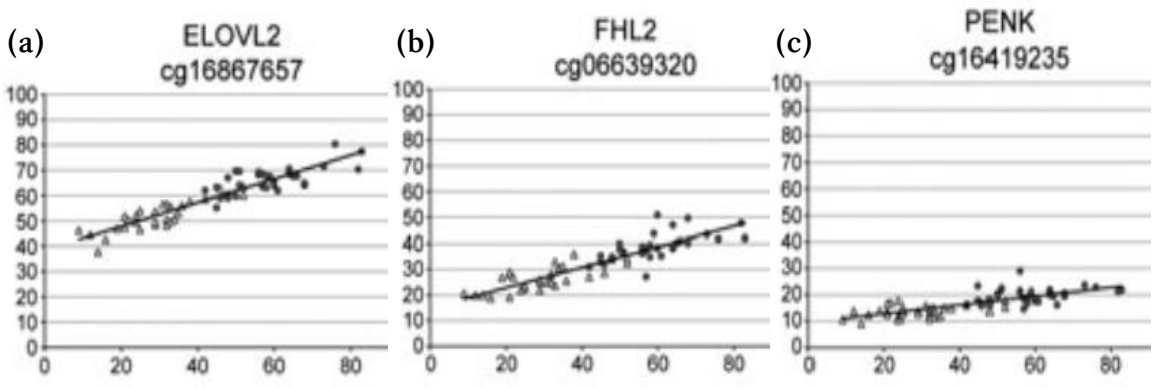


Figure 6-3. DNA methylation of selected CpG sites in (a) ELOVL2, (b) FHL2 and (c) PENK genes with respect to age (Garagnani *et al.*, 2012)
Circles represent methylation values in mothers while triangles the ones in offspring.

6.2.2 Effect of various environmental factors

It is known that similar to other physiological biochemical modifications DNA methylation is a reversible biological signal (Ramchandani *et al.*, 1999) that responds to various internal and external stimuli. Most of the observed age-associated DNA methylation profiles above are a result of various environmental factors including susceptibility to diseases and lifestyle. This phenomenon can be observed in monozygotic twins since older twin pairs demonstrate greater DNA methylation differences than younger ones (Fraga *et al.*, 2005) [Figure 1-3]. While environment during early embryogenesis may cause extensive, soma-wide DNA methylation modifications that can lead to fatal programing of adult disorders, environment later during life is more likely to induce less extensive, tissue-specific changes bringing about tissue-specific carcinogenesis (Lee & Pausova, 2013).

As an example, in a locus-by-locus analysis of exposure-related methylation, researchers identified 24 asbestos-related, 30 drinking-related and 138 smoking-related differentially methylated CpG sites in pleural tissues, blood and lung tissues respectively (Christensen *et al.*, 2009). Furthermore, chronic sun exposure has been reported to alter the methylation status of various keratin genes in human skin (Gronniger *et al.*, 2010). Additionally, following acute or chronic high dose (5 Gy) exposure of ionising radiation, researchers studied DNA methylation in liver, spleen and lung tissues of mice and found sex-, tissue- and dose-dependent radiation-induced DNA methylation changes (Pogribny *et al.*, 2004).

6.3 Age prediction models

From a forensic perspective, it would be very advantageous to be able to translate observed age-associated DNA methylation differences in a way that the chronological age of an individual is revealed. Even though building age prediction models using DNA methylation profiles is a relatively new field, there are a few studies that have focused on forensically relevant tissues, such as blood, saliva, and brain.

6.3.1 Blood

Hannum and his team performed one of the largest and highest-resolution genome-wide methylation studies to date using the whole blood of 656 individuals aged 19-101 years old (Hannum *et al.*, 2013). They obtained methylome-wide profiles by measuring the methylation status of more than 450,000 CpG sites with the Illumina Infinium Human Methylation 450K BeadChip assay. Remarkably, it was revealed that ~15% of the tested CpGs showed significant associations between their methylation levels and age. Applying a penalised multivariate regression method, the authors built a quantitative model using 71 highly age-predictive markers with a correlation between true and predicted age of 0.96 and an average error of 3.9 years [Figure 6-4a]. Nearly all of the markers applied in the model lay within or near genes with a known link to age-related conditions including DNA damage, cancer and Alzheimer's disease. As an example, there are two CpG sites that are located within the somatostatin gene (SST), which is known to play a key role in the regulation of the endocrine and nervous systems (Yacubova & Komuro, 2002). Given that the associations between ageing and prolonged life with metabolic activity are well established (Jumpertz *et al.*, 2011), it was expected that most of the methylation markers are connected with metabolism.

Interestingly they noticed that the methylome of men appeared to age approximately 4% faster than that of women [Figure 6-4b] and there were always a few individuals that appeared to be ageing faster or slower than what the model would predict. The model was also validated using an independent cohort of 174 samples and the same methodology; predictions were still highly accurate, with an error of 4.9 years (correlation of 91%) (Hannum *et al.*, 2013). Furthermore, the model was able to distinguish between very young and old individuals even when samples were processed following a different methodological approach (Heyn *et al.*, 2012).

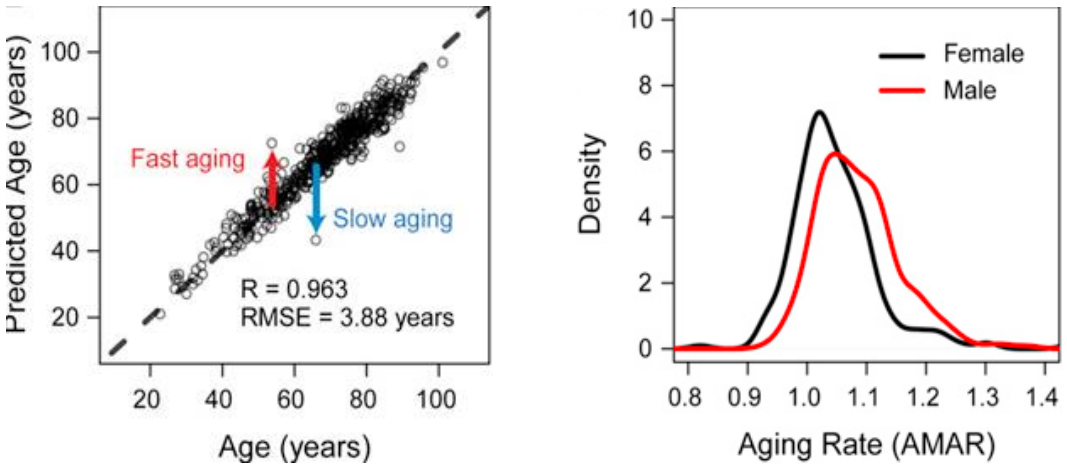


Figure 6-4. Age model predictions using 71 CpG sites in blood (Hannum *et al.*, 2013)

Predicted vs. chronological age for all individuals based on the ageing model. The red arrow indicates an individual that the predicted age was much higher than the actual age (fast ageing individual); similarly, the blue arrow shows an individual whom the model predicts as much younger (slow ageing individual).

From a forensic perspective, to analyse samples with this methodology would require large amounts of DNA which are usually not available. In an attempt to narrow down the number of age-associated markers needed for accurate prediction, Weidner *et al* performed a comprehensive analysis of methylation profiles and found that the methylation levels of only three CpGs – located in the integrin, alpha 2b (*ITGA2B*), aspartoacylase (*ASPA*) and phosphodiesterase 4C, cAMP specific (*PDE4C*) genes – were enough to create an epigenetic-ageing-signature (Weidner *et al.*, 2014). They developed a bisulphite Pyrosequencing® protocol after testing 82 whole blood samples that allowed age prediction with a mean absolute deviation (MAD) from chronological age of 5.4 years (RMSE=7.2 years) [Figure 6-5a]. The model was further validated using 69 blood samples and the observed predictions correlated with age even better (MAD=4.5 years, RMSE=5.6 years) [Figure 6-5b].

6.3.2 Saliva

Apart from blood, buccal epithelium has also been used for age-associated methylation analysis. Bocklandt *et al* performed a genome-wide methylation analysis using saliva samples of 34 pairs of male identical twins (21-55 years old) using Illumina's Human Methylation 27K microarray (Bocklandt *et al.*, 2011). The authors identified a total of 88 novel loci that were significantly correlated with age (absolute correlation values greater than 0.57) and were located near genes involved in cardiovascular, neurological or genetic diseases. Using only three CpG sites – located in the Edar

associated death domain (*EDARADD*), target of myb1 (chicken)-like 1 (*TOM1L1*), and neuronal pentraxin II (*NPTX2*) genes – they built a regression model that was linear with age over a range of five decades and explained 73% of the variance (MAD=5.2 years) [Figure 6-6].

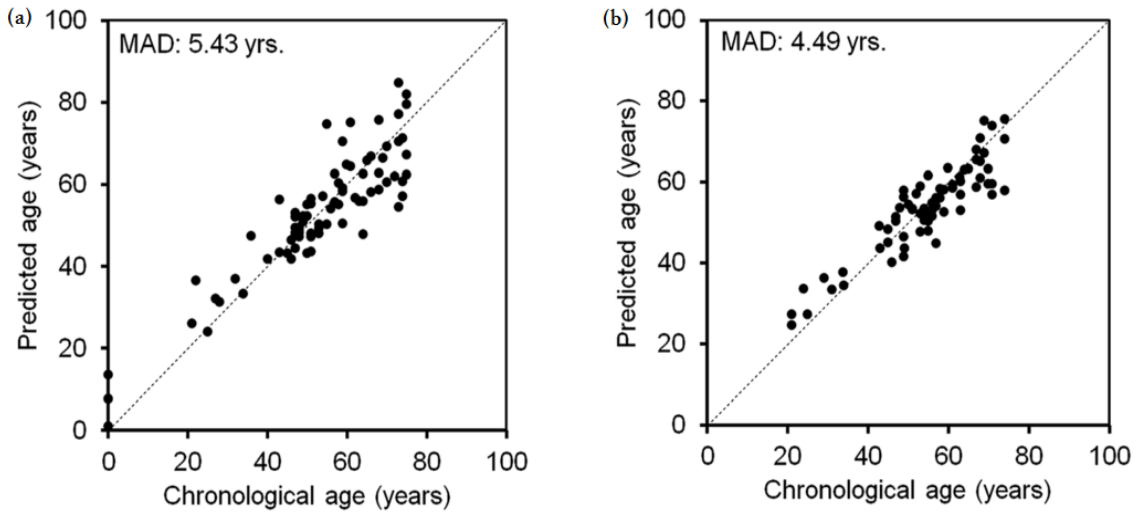


Figure 6-5. Age prediction model using three CpG sites in blood (Weidner *et al.*, 2014)
(a) Multivariate ageing model based on three CpG sites belonging to the genes *ITGA2B*, *ASPA* and *PDE4C* that allowed for accurate age prediction (MAD=5.4 years), (b) Slightly higher precision was obtained during validation with an independent set of 69 blood samples (MAD=4.5 years).

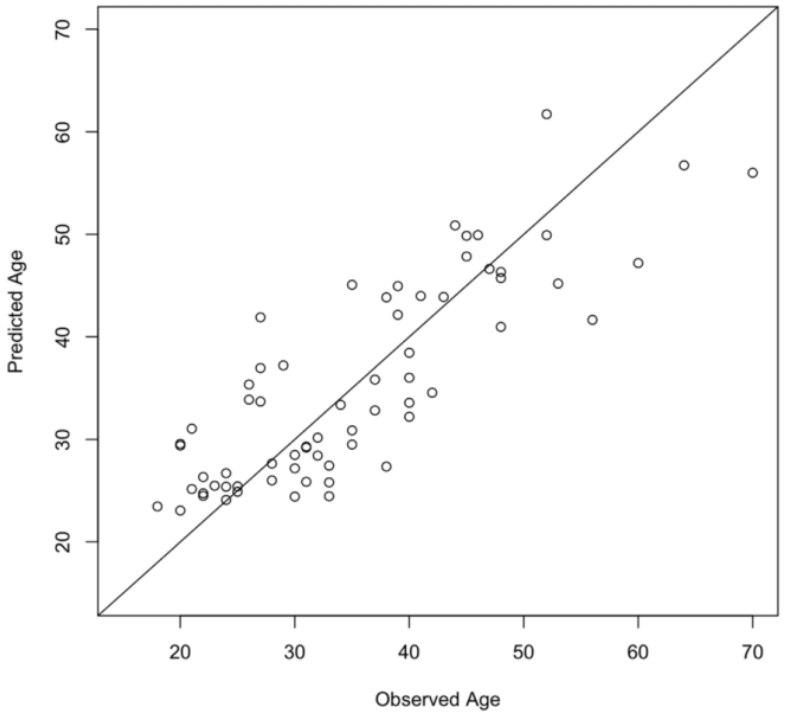


Figure 6-6. Predicted vs. observed age using a leave-one-out model in saliva (Bocklandt *et al.*, 2011)
Age prediction model by multivariate regression analysis showing high correlation between predicted and observed ages (r=0.83, MAD=5.2 years, n=66)

6.3.3 Multi-tissue

As previously mentioned, each tissue or body fluid shows a different age-associated DNA methylation pattern. It would be very useful to develop an age prediction test that could be applied in all human tissues and Kock and Wagner have gone some way towards this. They analysed several publicly available DNA methylation datasets that examined more than 27,000 CpG sites in 13 different cell types (Koch & Wagner, 2011). Initially they identified 431 age-associated hypermethylated, and 25 age-associated hypomethylated CpG sites. Next, they chose a subset of five markers – located in the tripartite motif containing 58 (*TRIM58*), potassium voltage-gated channel, KQT-like subfamily, member 1 downstream neighbour (non-protein coding) (*KCNQ1DN*), *NPTX2*, BIRC4-binding Protein (*BIRC4BP*) and glutamate receptor, ionotropic, AMPA 2 (*GRIA2*) genes – to be integrated into their epigenetic-ageing-signature test. Interestingly, one of these CpG sites (*NPTX2*) is also included in the previous study by Bocklandt and his colleagues. Based on these five CpG sites, their predictions had an average precision of ±9.3 years.

Predicting age across a broad spectrum of human tissues and cell types appears to be a very challenging task; using a much larger dataset could, however, potentially overcome the difficulties and provide a more accurate age prediction. Horvath developed a multi-tissue predictor of age by employing ~8,000 samples from publicly available Illumina DNA methylation array datasets (both 27K and 450K) comprised of 51 healthy tissues and cell types (Horvath, 2013). The author used a regression analysis that avoids over-fitting of the model (penalised regression with elastic net analysis) and selected 353 CpGs, named as the 'epigenetic clock'. Across all test data, the age correlation was 0.96 with an error of 3.6 years, which appears very promising. Individual models according to tissue type were also built to improve accuracy (Horvath, 2013). The age predictor worked well even in heterogeneous tissues such as whole blood, buccal epithelium, uterine cervix and saliva, but in sperm the methylated-DNA age was significantly lower than the chronological age of the donor.

6.4 Conclusion

As shown in this chapter, accurate molecular age estimation from different types of biological material is not a simple task as ageing is biologically complex. However, age-associated DNA methylation profiling is very promising and future research has the potential to influence our understanding of ageing and allow for more accurate predictions. Either considering healthy or diseased samples, the employed methodologies that analyse thousands of CpG sites at once usually require large amounts of intact DNA, which is often not available in forensic samples. Therefore, it is important to further investigate previously reported age-associated CpG sites in an attempt to develop assays that could be applicable in the forensic setting.

7 Chronological age prediction in blood using age-associated CpG sites

7.1 Introduction

As shown in Chapter 6, there have been various chemical- or biological-based methods that have been proposed in the literature; all of which suffer from limitations. In forensic science, accurate biological age estimation is not an easy task since specimens could be of different types of biological material and are often minute or degraded. Recent research shows that age-associated DNA methylation profiling could be a good alternative to current methods but further research is needed to allow for more accurate predictions. As described in section 6.3, proposed age prediction models are based on a large set of CpG sites analysed with genome-wide methylation analysis protocols meaning that large amounts of intact DNA are required. Therefore, it is important to assess the significance of each of the reported CpG sites and select the 'best' for further validation. Consequently, the purpose of this chapter was to further investigate previously reported age-associated CpG sites in an attempt to develop assays that could be applicable in a forensic setting where samples are often degraded. An ideal forensic age prediction method should require minimal starting DNA material and analyse as few markers as possible without at the same time sacrificing prediction accuracy and reproducibility.

So far, research has focused on predicting biological age in various tissues since the 'calculated' degree of ageing could act as a predictor in several age-related diseases. On the other hand, from an investigative perspective where information such as disease state or environmental exposure is rarely available, it is important that the CpG sites selected for further analysis are chosen only from control DNA samples. Also, since the focus of this project is the age prediction of living individuals and blood is the most common body fluid found at crime scenes, it was decided that efforts would be initially focused on creating an age prediction model that could be applied in bloodstains.

Again since forensic samples are very often of low quantity and quality, applying genome-wide DNA methylation analysis techniques would not be appropriate. Thus, it is important to develop suitable and sensitive assays that can accurately quantify the methylation levels of specific CpG sites. Since bisulphite Pyrosequencing® was successfully applied in identifying the tissue source of a sample as shown in Chapter 5, it was believed that it could serve as a potential method for the application of age

estimation as well. As demonstrated in various ageing studies, bisulphite Pyrosequencing® has been effectively applied as a confirmation method to validate age-associated CpG sites reported in array-based results (Alisch *et al.*, 2012; Bocklandt *et al.*, 2011; Christensen *et al.*, 2009; Weidner *et al.*, 2014).

7.1.1 Aim and Objectives

The aim of this study was to investigate selected age-associated CpG sites and assess the possibility of using a small number of DNA methylation markers to accurately predict chronological age.

In order to meet the aim, two different approaches were followed:

- Firstly, a set of reported age-associated CpG sites in blood was selected and bisulphite Pyrosequencing® assays were designed to further validate their methylation levels. Once these assays were developed and optimised, whole blood samples of known age were analysed. Age predictions were then performed and the accuracy of the designed model was assessed.
- Secondly, a set of reported age-associated CpG sites in various tissues was selected and publicly available genome-wide DNA methylation data in blood were collected regarding these markers. The resulting dataset was analysed by various analysis methods in order to create an accurate age prediction model. The model was further validated using diseased blood samples or samples of different tissue origin.

7.2 Experimental

7.2.1 Approach 1

7.2.1.1 Blood samples

Whole peripheral blood was collected from 90 healthy volunteers (53 males and 37 females from various ethnic backgrounds) aged 1 week to 85 years old (mean age 35.1±17.4 years) as described in section 2.1 [Figure 7-1]. Blood samples were stored at 2-8 ºC either as a liquid (200-1,000 µl) or spotted onto a Whatman classic card.

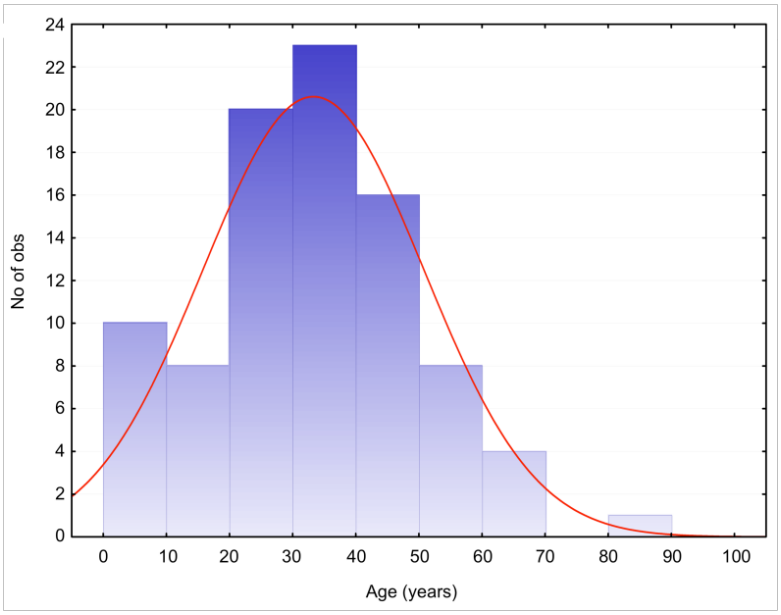


Figure 7-1. Age distribution within the blood sample database (n=90)
The histogram illustrates the age distribution of all individuals participated in the study, which resembles the normal (or Guassian) distribution.

7.2.1.2 Selection of age-associated CpG sites in blood

As previously mentioned, all age prediction models found in the literature show good prediction accuracy, however, not without limitations. For this part of the project, the study by Hannum *et al.* (2013) was selected as the most appropriate to investigate further for a number of reasons including the number of samples and sites used. Considering that this study was the first to investigate more than 450,000 CpG sites (high resolution), analyse more than 650 samples (large dataset) and build an age prediction model with less than 4 years of error (high accuracy), it was concluded that it was the most appropriate dataset to validate methylation markers from.

The authors categorised the group of 71 markers included in the model using an age coefficient, ranging from -22.7 to 28. This coefficient represents the link between the particular methylation marker and ageing and can be either positive or negative. A positive coefficient indicates that methylation increases with age, while a negative coefficient is a sign of a decrease in methylation with time. As expected, the more positive or negative the coefficient is, the higher the importance of the CpG site is in the age model. When selecting the set of CpG sites for further investigation, the age coefficient was the only parameter available regarding the relationship of the markers with ageing so it was the sole basis for marker selection. Ten of these CpG sites that showed the 'strongest' positive or negative correlation with age were chosen for the development of Pyrosequencing® assays [Table 7-1]. All selected sites are located near human genes across six chromosomes and half of them belong to a 'CpG island' too (300-3000 bp long, >55% GC content). Also, most of them (7/10) are positively associated with age (positive age coefficient).

7.2.1.3 Bisulphite Pyrosequencing® assay design

Following the guidelines regarding assay design mentioned in sections 2.2.4.1 and 2.2.6.1.1, bisulphite Pyrosequencing® protocols were designed using the BiSearch primer design tool. Two of the selected CpG sites were in close proximity (cg11067179 and cg22213242) so a common assay was designed resulting in nine assays in total (AGE1-9) [Tables 7-2 & 7-3]. Each assay includes a 10X PCR primer set (forward and reverse) as well as a 10X sequencing primer for use in the Pyrosequencing® reaction (all reverse primers are 5' biotin-labelled). Sequencing primers were designed to bind in a location close to the CpG site in question avoiding (where possible) the presence of another CpG site. Finally, all primers were designed to include converted cytosines in their sequence so they only bind to bisulphite-converted DNA.

7.2.1.4 Sample analysis

Genomic DNA from 4 μl or a 1.2 mm dried spot of each blood sample was extracted using the manually performed Chelex method as described in section 2.2.1.1. A total of 10 μl of extracted DNA (~10 ng) or 100 ng of each DNA methylation standard were then treated with sodium bisulphite using the MethylEdge™ Bisulphite

Conversion System (Promega) as described in section 2.2.3.3; bisulphite-converted DNA was eluted in 15 µl of elution buffer and stored at 2-8 °C for up to one week. All samples were amplified in duplicate using the ZymoTaq™ premix (Zymo Research) and following the optimised conditions of each PCR assay. Briefly, each PCR reaction consisted of 12.5 µl of ZymoTaq PreMix, 1 µl of 25 mM $MgCl_2$ for a final concentration of 2.75 mM (since the ZymoTaq™ Premix also contains 1.75 mM $MgCl_2$), 1 µl of each PCR primer (for a final concentration of 0.4 µM), 1 µl of bisulphite DNA template and 8.5 µl of nuclease-free water, for a total reaction volume of 25 µl. The thermocycling program used was: 95 °C for 10 minutes, followed by 45 cycles of 94 °C for 30 seconds, T_m (AGE1/3 - 50 °C, AGE6 - 51 °C, AGE2/5 - 53 °C, AGE4/7/8/9 - 55 °C) for 30 seconds, 72 °C for 30 seconds, and a final extension step of 72 °C for 7 minutes. Following amplification, the quality of PCR products was assessed on a 2% agarose gel as described in section 2.2.5.1. Lastly, using 10 µl of biotinylated PCR products Pyrosequencing® was performed as described in section 2.2.6.1 and DNA methylation values were obtained via the PyroMark CpG software (QIAGEN).

Table 7-1. Selected age-associated CpG sites for validation via bisulphite Pyrosequencing®
Genomic information of each marker including chromosomal position, associated gene(s) and CpG island together with age coefficients.

Marker	Chromosomal position	Genes	CpG island	Coefficient
cg05442902	22:21,369,010	LZTR1, MIR649, P2RX6, SLC7A4, THAP7	No	-22.7
cg09651136	15:72,525,012	PARP6, PKM2	No	-15.8
cg20822990	1:17,338,766	ATP13A2, SDHB	No	-15.7
cg11067179	11:66,083,541	CD248, RIN1, TMEM151A	No	14.7
cg21139312	17:55,663,225	MSI2	No	17.1
cg20426994	7:130,418,324	KLF14	Yes	19.1
cg14692377	17:28,562,685	BLMH, SLC6A4, SNORD63.3	Yes	19.1
cg22213242	11:66,083,573	CD248, RIN1, TMEM151A	Yes	23.7
cg08097417	7:130,419,133	KLF14	Yes	27.3
cg03399905	15:79,576,060	ANKRD34C	Yes	28

Table 7-2. Designed bisulphite PCR assays

Essential information regarding the designed bisulphite PCR assays is shown, such as the score assigned by the primer design software (the lower the score the more efficient amplification), the primer sequences and length (F for forward and R for reverse), the % G and C content, the melting temperature (T_m), the number of converted cytosines included in the primer sequence (highlighted in red) as well as the length of the final PCR product.

CpG site	Assay	Score	Primer Sequence (5' → 3')	Length (bp)	%GC	T_m (C°)	Converted Cs	PCR Product (bp)
cg03399905	AGE1	39.54	F TATAGATATTGGTAATAATG	20	20.0	49.8	5	170
			R TAACTACCCTAAACTAAA	18	27.8	53.4	5	
cg08097417	AGE2	28.35	F GGTTAAGTTATGTTTAATAGT	21	23.8	55.0	3	198
			R AAAACTTTCTAAAACTCC	18	27.8	54.4	3	
cg11067179 cg22213242	AGE3	38.17	F GTTATTAGTTTTTAGTTTATG	21	19.0	53.4	9	171
			R AACTACAACTACCACTAT	18	33.3	52.8	4	
cg05442902	AGE4	25.20	F TTTTGTGTTTTTTAGTTATTTTG	23	17.4	57.4	6	120
			R AACTAACCCTTACAAATTTC	20	30.0	56.9	4	
cg20426994	AGE5	21.05	F AATAGGTTTTGGTGTAGTT	19	31.6	56.0	4	138
			R CAACCTCTAATAAATTCTCT	20	30.0	54.6	6	
cg14692377	AGE6	37.93	F GATTTATTTGTTAGGTTG	18	27.8	51.8	8	148
			R ATCAAACCATATAAAAAC	18	22.2	51.1	6	
cg21139312	AGE7	13.98	F AGAAAGTTTTGAGTTGAGAA	20	30.0	56.5	3	194
			R CCAACAAAAAATACCAAAAC	20	30.0	57.3	7	
cg09651136	AGE8	13.05	F AAATTAAGAATAGTGGTGGAT	21	28.6	56.5	3	95
			R CTACACCCAAACAATTTAAAACT	23	30.4	60.1	3	
cg20822990	AGE9	19.99	F GTTTGTTTTATAGAGAATTGTTG	23	26.1	57.1	7	183
			R CTCTTTTTACCCATACTAAAAT	22	27.3	57.4	5	

Table 7-3. Pyrosequencing® DNA methylation assays

The table includes details of the sequenced chromosomal locations, the sequencing primer sequences as well as the nucleotide dispensation order. The ten CpG sites in question are highlighted in grey, bisulphite-conversion controls in bold while dead injections are underlined.

Assays	Chromosomal location	Sequencing primer (5' → 3')	Sequence to be analysed (5' → 3')	Dispensation order
AGE1	15:79,576,059-068	TAGATATTGGTAATAATGG	ACGCTGGCAC	GATCGTCAGTCATC
AGE2	7:130,419,112-135	TAATAGTTTTAGAAATTATTTTG	TCTCCGCGTTCTTTCTTCTGCCGG	ATCGTCGTCAGTCG
AGE3	11:66,083,539-575	ATATATATTTGTTGGTATAT	ACCGGCAATCTGGCACTCATCTGTGTCCA CACAGCGG	GATCGTCATCGATCATCATCGTGTCATCAT CATGTCG
AGE4	22: 21,369,009-021	TTGTATGTTTTGGTTTTTGTAT	ACGCTGCTCCCTG	GATCGTCGATCG
AGE5	7:130,418,302-328	TTTGGTAGTAGGTGTGATAG	ACCTCCTCCGGGGCGCCTGATCCGCGG	TATCGTCGTCGAGTCGTCG
AGE6	17: 28,562,680-695	GGTTGGGTAGGCGGGTTGG	CCTCGCGCCCTCGAGG	ATCGTCGATCGAG
AGE7	17:55,663,213-239	TTTTGTGGTTTTTAGAGTAGAT	GCGGAGGCGGCACGTCCTCGTGCCCTT	AGTCGAGTCGTCATCGATCGTGTC
AGE8	15:72,525,009-018	ATATTGGGATGGTAAGGATTT	GGCCGATGCT	AGTCGATAGTC
AGE9	1:17,338,763-777	TATATAAGAAAATGTTTGTTAATT	GGGCGTGGTGGCGCA	AGTCGTGATGTCGTCA

7.2.2 Approach 2

7.2.2.1 Publicly available DNA methylation data

Genome-wide profiling has led to more comprehensive understanding of gene regulation epigenetic mechanisms. Illumina's Human Methylation BeadChip technology is one of the most commonly used genome-wide methylation platforms that allows for simultaneous measurement of the methylation status of over 27,000 (27K chip) or 450,000 (450K chip) CpG sites in the genome at single nucleotide resolution (Bibikova *et al.*, 2011). Thousands of samples have been assayed using this platform in the literature and researchers have made some of these genome-wide methylation data available in online databases. For the purpose of this study, due to restrictions in time and resources it was decided that online datasets would be selected that could offer a large pool of samples for age prediction.

Prior to collection of the DNA methylation data for the selected age-associated CpG sites, these genome-wide DNA methylation datasets underwent a quality control analysis. Experimental biases on these genome-wide platforms are common and efforts should be made to eliminate these prior to data analysis. In this study, these analyses took into account, where available, the chip used, the order of each sample on the chip and the dataset (factor variables). Ideally, the reported bisulphite conversion rates should have been used too, however these data were not available for any dataset included in this study. Using a linear regression model, the normalised methylation beta values on these factor variables (resulting residuals) were created and used for age prediction analysis.

Whole blood

In his study, Horvath (2013) used available methylation data gathered from whole blood, cord blood or individual blood cell types such as peripheral blood mononuclear cells (PBMCs) or CD4+ cells. Furthermore, it was assumed that certain disease statuses such as schizophrenia had negligible effect on age relationships; therefore, diseased samples were also included in the model. For the purpose of this project, which was predicting age from blood stains, it was decided that only data obtained from whole peripheral blood of volunteers used as controls was suitable.

Datasets included in Horvath's study together with recently published data that included age as one of their parameters comprised the final sample database of this study as shown in Table 7-4. The database included a total of 1,156 whole blood samples collected from individuals aged between 2-90 years old (mean age=44) and from various ethnic backgrounds as part of seven genome-wide DNA methylation studies. Samples were carefully collected so that there was an equal representation of samples for all age groups aiming for ~100-150 samples per decade, which was quite challenging [Figure 7-2]. Since the gathered samples were either healthy control samples in studies investigating DNA methylation changes of various diseases (usually above 40 years old) or were part of studies investigating epigenetic effects of ageing (usually either very young or very old), collecting sufficient samples of 'middle' age (especially 30-40 years old) was quite difficult. Also, even though the database included both genders in equal amounts (597 females and 559 males), there was an uneven gender distribution within the age groups due to the selected studies' design [Figure 7-2]. However, it was concluded that this should not affect age prediction since none of the sex-specific differentially methylated CpG sites reported in the literature following analysis with the Illumina's 27K platform were included in the group of the selected 45 markers (Chen *et al.*, 2011).

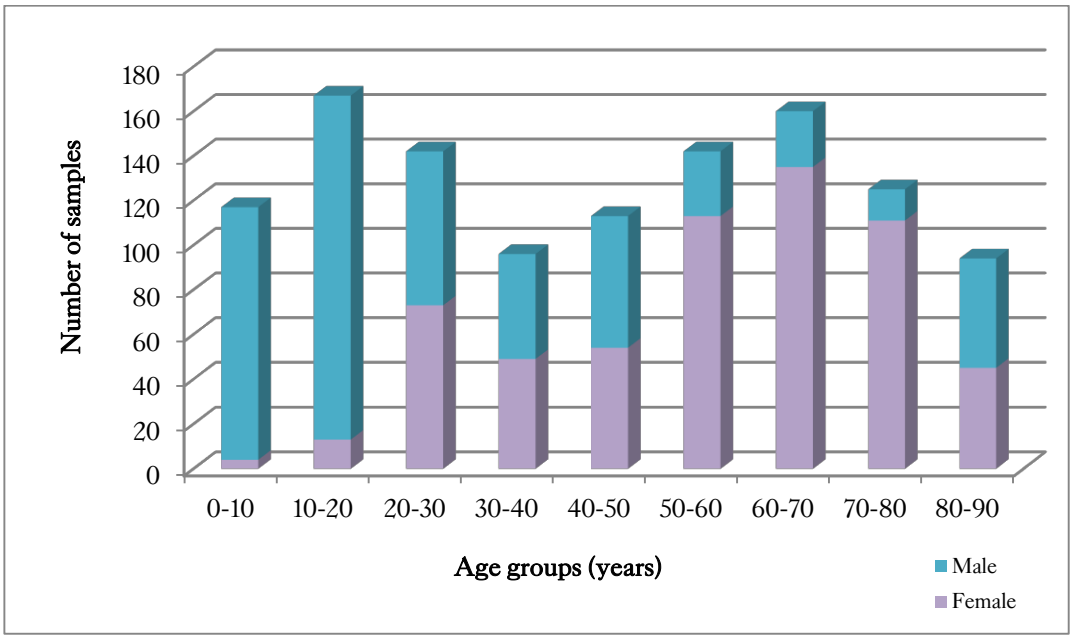


Figure 7-2. Age distribution of samples used in the age prediction model (n=1,156)

Table 7-4. DNA methylation datasets from various healthy tissues
The table provides an overview of the selected DNA methylation data. Information regarding the sample size, gender (F/M), age range, mean age and applied Illumina platform are included in the table. All data are publicly available and can be assessed using the Accession No in the Gene Expression Omnibus (GEO) database.

Tissue	Set	Samples	♀/♂	Age range (mean) (years)	Platform	Accession No
Whole blood	1	235	0/235	4-18 (10)	27K	GSE27097
	2	24	12/12	2-35 (14)	27K	GSE23638
	3	170	77/93	32-90 (65)	450K	GSE40279
	4	385	187/198	16-88 (40)	27K	GSE41037
	5	33	12/21	18-65 (29)	450K	GSE41169
	6	91	91/0	49-74 (63)	27K	GSE20236
	7	218	218/0	52-78 (65)	27K	GSE19711
	8	40	40/0	47-77 (62)	450K	GSE53128
	9	66	66/0	33-68 (56)	27K	GSE58045
Saliva	10	69	0/69	21-55 (35)	27K	GSE28746
	11	196	54/142	21-55 (32)	27K	GSE34035
Buccal cells	12	5	3/2	15	27K	GSE25892
	13	4	2/2	4-5	450K	GSE42409
	14	10	5/5	28-47 (35)	450K	GSE50586
Skin	15	15	15/0	6-73 (34)	27K	GSE22595
	16	30	30/0	18-59 (39)	27K	E-MTAB-625
	17	8	8/0	43-62 (49)	27K	GSE29661
Cervix	18	167	167/0	19-69 (29)	27K	GSE30758
Muscle	19	11	6/5	53-78 (68)	27K	GSE38291
	20	51	0/51	21/77 (50)	27K	GSE49908

In more detail, set 1 includes methylation profiles of 235 healthy boys aged 4-18 years as part of a study into paediatric age-associated DNA methylation (Alisch *et al.*, 2012), while set 2 consists of 12 female and 12 male samples in an attempt to investigate autosomal sex difference in the DNA methylome (Chen *et al.*, 2011). Set 3 includes 170 methylation profiles of individuals across a large age range as part of a study into human ageing rates (Hannum *et al.*, 2013), whereas set 4 and 5 are comprised of 385 and 33 healthy control Dutch subjects of a study investigating DNA methylation changes in schizophrenia patients using Illumina's 27K and 450K arrays respectively

(Horvath *et al.*, 2012). Lastly, set 6 includes 91 methylation profiles of females aged 49-74 years as part of a study investigating age-associated hypermethylation (Rakyan *et al.*, 2010) and set 7 is comprised of 218 methylation profiles of healthy females aged 52-78 years participating in an ovarian cancer population study (Teschendorff *et al.*, 2010).

In order to investigate environmental influences on age prediction, the developed model was further validated using an independent cohort of healthy blood samples comprising of 53 female monozygotic twin pairs collected from two genome-wide studies [Table 7-4]. Set 8 includes 40 methylation profiles of females aged 47-77 years that were used to investigate differential methylation patterns involved in pain sensitivity (Bell *et al.*, 2014), while set 9 is comprised of 66 female individuals aged 33-68 years participating in an age-associated methylation study (Bell *et al.*, 2012).

Other tissues

Since the set of 353 age-associated CpG sites included in the 'epigenetic clock' was suggested to be robust across various tissues, it was believed that the results of the developed model based on blood could potentially be also applicable in other tissues. Tissues including saliva, skin, cervix and muscle were chosen since they could be of forensic value. Sperm was not included in this study since there was no significant correlation between the predicted and the chronological age of the donors (Horvath, 2013). The main challenge was that compared to blood there were not as many studies available that had used one of the Illumina platforms and had information regarding the volunteers' age at the same time. This was considered as one of the key limitations when aiming to build a model in tissues other than blood. However, a final set of 566 samples for the five selected tissues was assembled using methylation data from eight different genome-wide methylation studies [Table 7-4].

In more depth, a total of 265 saliva samples from individuals aged 21 to 55 years were collected from two different studies. Set 10 includes 69 samples as part of an age-associated genome-wide epigenetic analysis (Bocklandt *et al.*, 2011), while set 11 is comprised of 196 individuals aged 21 to 55 years that volunteered in a study on the influence of sex on genome-wide methylation (Liu *et al.*, 2010). As for buccal cells, three studies were chosen including five adolescents (15 years old) recruited in a study

of early developmental adversity (set 12) (Essex *et al.*, 2013), four children aged 4-5 years old (set 13) (Price *et al.*, 2013) and also, ten individuals participated in a study investigating DNA methylation patterns in Down syndrome (set 14) (Jones *et al.*, 2013). Moreover, 53 skin samples from females aged 6 to 73 years were collected from three studies including a study investigating age-associated DNA methylation changes in dermal fibroblasts (set 15) (Koch *et al.*, 2011), an epigenetic analysis of ethnic variations (set 16) (Winnefeld *et al.*, 2012) as well as a study tracking cellular ageing (set 17) (Koch *et al.*, 2012). Also, cervix samples were collected from 167 females (19-69 years old) from a study investigating the association between epigenetic variability and risk of future morphological transformation (set 18) (Zhuang *et al.*, 2012). Lastly, 62 skeletal muscle samples from individuals aged 21-78 years were collected from two different studies, both invesigating methylation differences between muscle and other tissues (set 19 and 20) (Day *et al.*, 2013; Ribel-Madsen *et al.*, 2012).

Diseased samples

According to Horvath (2013), the correlation between the observed and expected age in cancer/affected tissues was generally weak as there was evidence of significant age acceleration in most patients included in his study (n=5,826). However, since there is usually no information regarding possible disease status in a forensic blood stain of unknown origin, it is important that the proposed age prediction model can be universally applied. To assess potential variability in age prediction, a data set including blood samples from a total of 1,011 (577 females and 434 males) individuals aged 17 to 91 years suffering from various diseases and cancers analysed on Illumina's 27K or 450K platforms was analysed [Table 7-5].

7.2.2.2 Selection of multi-tissue age-associated CpG sites

The ability to accurately predict age regardless of the tissue would be very advantageous in forensic science where the identification of the tissue source of a sample is often challenging. Even if the purpose of this project was to identify age-associated CpG sites in blood, the ability to apply a potential model in other tissues would save both time and resources. In an attempt to select more robust age-associated differentially methylated markers, the study by Horvath (2013) was chosen for a number of reasons.

Table 7-5. DNA methylation datasets with various diseases

Disease	Samples (♀/♂)	Age range (years)	Study	Accession No
Type 1 diabetes	194 (98/96)	24-74	(Bell *et al.*, 2010)	GSE20067
Anaemia	28 (24/4)	23-85	(Day *et al.*, 2013)	GSE49904
Schizophrenia	385 (97/288)	17-86	(Horvath *et al.*, 2012)	GSE41037 GSE41169
Bone marrow disorders	77 (31/46)	28-90	(Perez *et al.*, 2013)	GSE42042
Ovarian cancer	262 (262/0)	49-91	(Teschendorff *et al.*, 2010)	GSE19711
Breast cancer	30 (30/0)	50-70	(Zhuang *et al.*, 2012)	GSE32396
	35 (35/0)	24-70	(Anjum *et al.*, 2014)	GSE57285

Mainly, the proposed multi-tissue model based on 353 CpG sites was built using the largest dataset published so far containing more than 8,000 samples gathered from dozens of genome-wide studies. The selected publicly available data were obtained from various laboratories worldwide employing either Illumina's 27K or 450K platform. Since the author decided to analyse only the common CpG sites between these platforms (21,369 in total), a larger dataset could be created. It was believed that this approach accounted for potential lab-to-lab variations or potential artefacts resulting in more dominant outcomes. Also, the ability of the test to predict age in whole blood with such high accuracy (mean absolute error=3.7 years, age correlation=0.95, test data) was a great advantage. The author categorised the group of 353 markers using a coefficient value (ranging from -1.719 to 3.067) that relates the CpG sites to a transformed version of age as well as a marginal (positive or negative) age relationship. In order to cover all potential correlation with age and maximise the change of selecting 'strong' markers, it was decided that a total of 45 CpG sites would be investigated. These include three sets, the first one containing the top 15 CpG sites displaying negative coefficients (-0.578 to -1.719), the second one containing the top 15 CpGs with positive coefficients (0.745 to 3.067) as well as a set with 15 CpG sites having coefficient values around zero (-0.002 to 0.006) [Table 7-6]. However, it should be noted that a future approach could include all 353 CpG sites. Their location was confirmed using the Ensembl genome browser; most of them are located across the genome within or near a gene. Interestingly, there is one common marker (cg05442902) with the previously selected set of CpG sites (Hannum *et al*, 2013).

Table 7-6. Selected 45 age-associated CpG sites from Horvath 2013

	CpG sites	Coefficient Value	Age Relationship	Gene
Set 1	cg16408394	-1.719	negative	RXRA
	cg25683012	-1.392	positive	BAZ2A
	cg19761273	-0.887	negative	CSNK1D
	cg27544190	-0.869	negative	C21orf63
	cg03588357	-0.859	positive	GPR68
	cg03286783	-0.855	negative	CASC4
	cg19273182	-0.755	positive	PAPOLG
	cg15703512	-0.731	negative	MGC50721
	cg01511567	-0.686	negative	SSRP1
	cg09441152	-0.683	negative	PQLC1
	cg02047577	-0.658	negative	UCKL1
	cg17338403	-0.645	negative	SLCO3A1
	cg07158339	-0.626	negative	FXN
	cg01873645	-0.604	positive	C9orf85
	cg05442902	-0.578	negative	P2RXL1
Set 2	CpG sites	Coefficient Value	Age Relationship	Gene
	cg04999691	-0.002	negative	C7orf29
	cg24450312	-0.002	positive	RASSF5
	cg04452713	-0.001	negative	DST
	cg22613010	-0.001	negative	CLCN2
	cg09646392	0.000	negative	TNFSF13B
	cg17274064	0.000	negative	ERG
	cg16984944	0.000	negative	TBC1D23
	cg00436603	0.000	negative	CYP2E1
	cg24126851	0.001	positive	DCHS1
	cg14723032	0.001	positive	PITPNM3
	cg06926735	0.001	negative	UBE2V1
	cg14308452	0.003	positive	MGC24975
	cg00374717	0.005	positive	ARSG
	cg07455279	0.006	positive	NDUFA3
	cg02085507	0.006	positive	TRIP10
Set 3	CpG sites	Coefficient Value	Age Relationship	Gene
	cg20692569	0.745	positive	FZD9
	cg04528819	0.772	positive	KLF14
	cg08370996	0.813	positive	NR2F2
	cg26297688	0.960	positive	C12orf23
	cg23092072	1.000	negative	AFF1
	cg04084157	1.034	positive	VGF
	cg01968178	1.169	positive	REEP1
	cg25505610	1.502	positive	hfl-B5
	cg06993413	1.503	positive	DPP8
	cg00864867	1.600	positive	PAWR
	cg22736354	1.719	positive	NHLRC1
	cg06493994	1.858	positive	SCGN
	cg02479575	1.875	positive	C19orf30
	cg16241714	2.552	positive	CEBPD
	cg14424579	3.067	positive	FLJ21839

7.3 Results

7.3.1 Investigation of 10 age-associated CpG sites via Pyrosequencing®

7.3.1.1 Optimisation of Pyrosequencing®-based assays

As already mentioned in Chapter 5, bisulphite-treated DNA is considered as one of the most challenging templates to amplify. Therefore it was necessary to optimise the PCR reactions for each marker to avoid mis-priming, primer self-annealing or non-specific PCR products. Each assay was optimised using an annealing temperature gradient, various concentrations of $MgCl_2$ and primer as well as different PCR cycling conditions. Figure 7-3 shows the successful amplification of the designed PCR products on a 2% agarose gel using a commercially available DNA methylation standard. As shown, even though PCR efficiency was higher for some assays, non-specific products were absent.

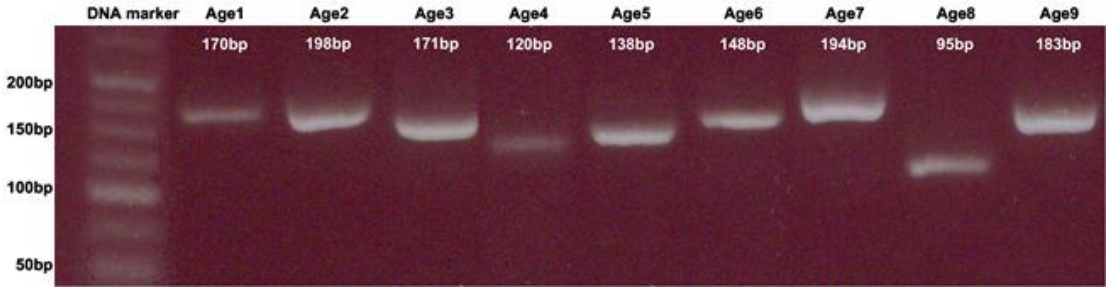
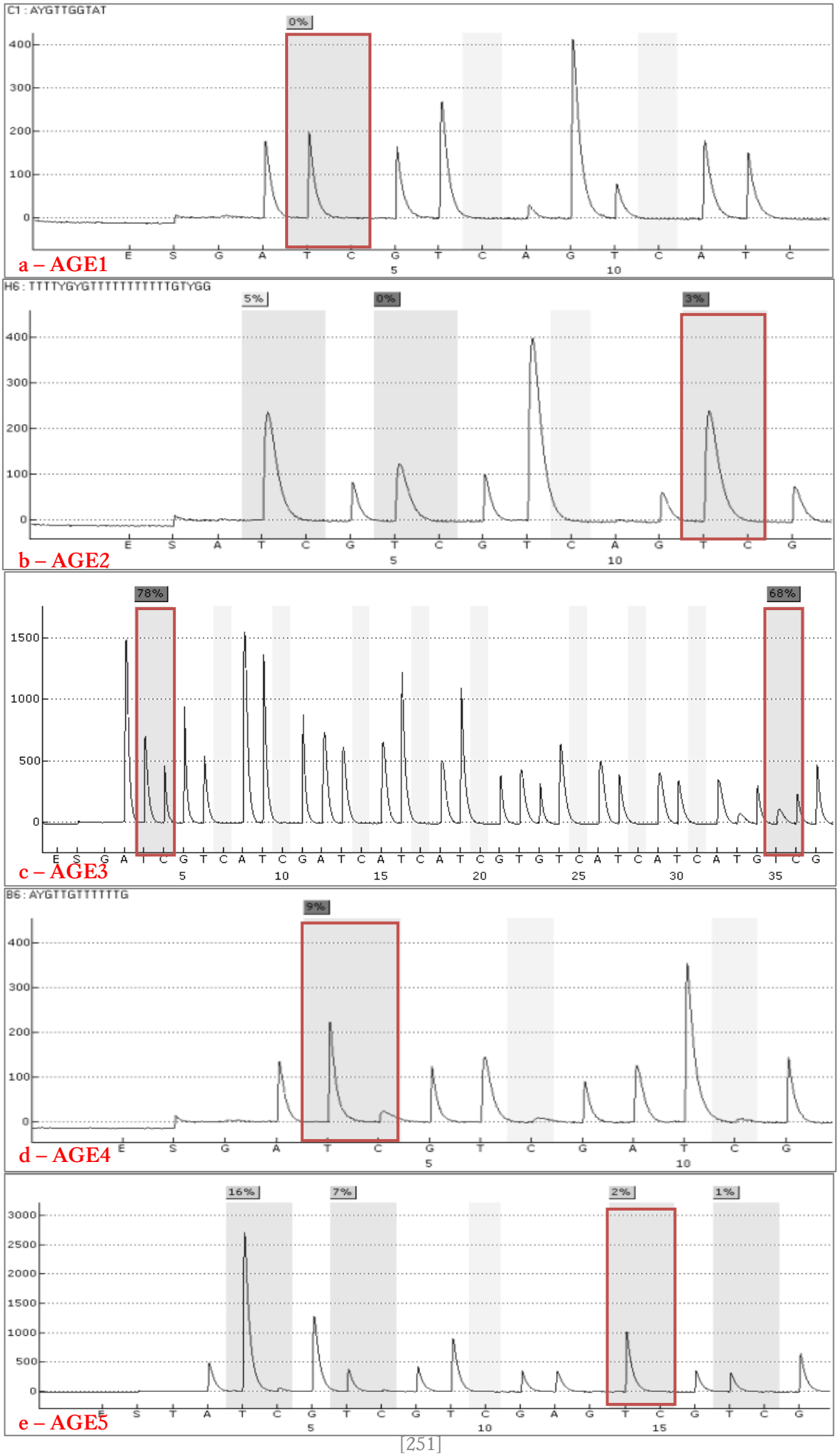


Figure 7-3. Final optimised PCR amplicons for assays AGE1-9
Successful amplification of a methylation DNA standard using all nine assays. Expected PCR lengths are shown on the top of the picture, together with the expected fragments of the DNA marker on the left. It is important to note that no non-specific amplification is present in any of the assays; however assays AGE1 and AGE4 seem to result in a weaker signal.

Optimised PCR products were then sequenced and methylation quantification was performed. Figure 7-4 provides example pyrograms™ obtained for one blood sample following analysis with all designed assays (AGE1-9). Bisulphite conversion rates were in most cases >98%; if bisulphite conversion rates were lower than 95%, the treatment with sodium bisulphite was repeated. Also, there was no chemiluminescence observed in the control injections suggesting that it was only the PCR product being sequenced. However, for certain assays the peaks were wider than expected indicating too much DNA input due to high PCR efficiency [Figure 7-4b]. Also, it was noted that occasionally the signal was decreased towards the end of the sequence probably due to incomplete incorporation of nucleotides [Figure 7-4c].


C1 : AYGTTGGTAT
0%
a – AGE1
H6 : TTTTYGYGTTTTTTTTTTTGTYGG
5%
0%
3%
b – AGE2
78%
68%
c – AGE3
B6 : AYGTTGTTTTTG
9%
d – AGE4
16%
7%
2%
1%
e – AGE5

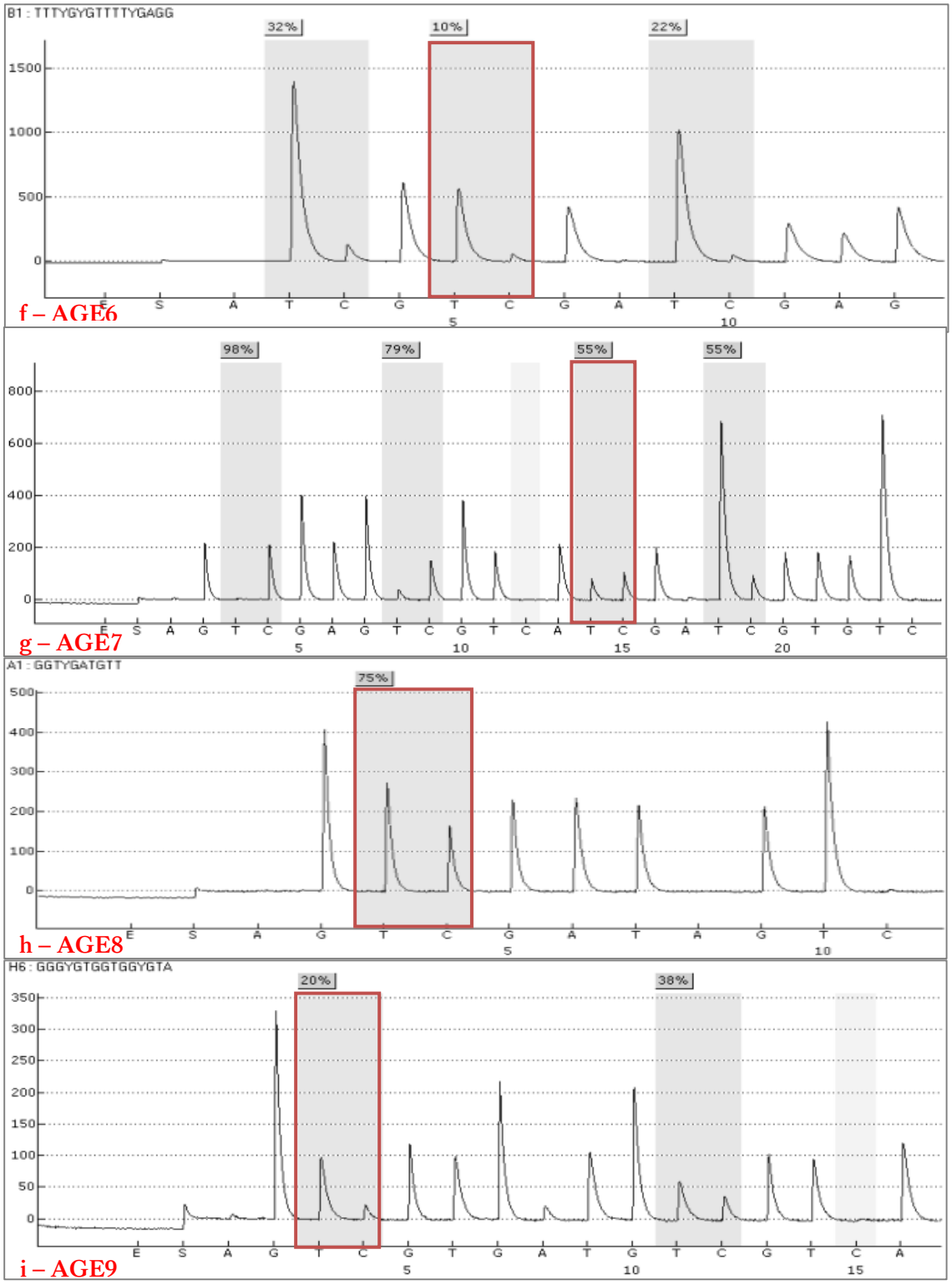


Figure 7-4. Example pyrograms™ of all designed pyrosequencing® assays AGE1-9

Although there are ten CpGs in question (highlighted in red blocks), in some assays adjacent CpGs were also quantified. Obtained methylation values are presented on the top, where the peak for 'T' corresponds to the unmethylated fraction (together with adjacent Ts or non-CpG converted Cs) while the peak for 'C' to the methylated fraction only. Light grey boxes indicate the position of built-in bisulphite-conversion controls.

7.3.1.2 Epigenetic drift in blood

Bisulphite-treated DNA samples obtained for 90 individuals (age range 1 week to 85 years old) were tested for all ten CpG sites following the experimental procedure explained in section 8.2.1.4. Only samples that passed the pyrosequencer's inbuilt quality controls in each assay were included in the analysis. The obtained individual methylation values were used to calculate the mean, standard deviation and methylation range for each assay [Table 7-7]. Some CpG sites showed higher variation than others. The range of methylation varied between 0.12 (cg03399905) and 0.99 (cg09651136), although for the latter marker the range would decrease to 0.59 if one single sample (0.01, 39 years old) was excluded.

Methylation fractions (zero to one) were then plotted against the actual age of each individual in order to investigate potential correlation between methylation levels and age. In general, no significant linear correlation ($p>0.05$) was observed for any of the ten CpG sites tested. As depicted in Figure 7-5, methylation levels for four markers cg03399905, cg05442902, cg21139312 and cg09651136 decrease with age; this is in accordance with the previously published negative age coefficient values for three of them (cg05442902, cg21139312 and cg09651136). Interestingly, during the marker selection process cg03399905 was chosen because it had the highest positive coefficient value at 28, yet methylation levels are shown to decrease with age from the data presented here. Also, there were three markers comprising cg08097417 and the two markers included in the AGE3 assay showing an increase in methylation with age. The rest did not seem to exhibit any significant distribution.

Table 7-7. Observed methylation variation for all tested CpG sites (10)

Markers	No of samples	Methylation Fraction			
		Minimum	Maximum	Range	Mean ± StDev
cg03399905	58	0	0.39	0.39	0.117 ± 0.125
cg08097417	72	0	0.12	0.12	0.031 ± 0.021
cg11067179	76	0.02	0.92	0.9	0.541 ± 0.201
cg22213242	74	0.26	0.71	0.45	0.552 ± 0.109
cg05442902	79	0.06	0.39	0.33	0.136 ± 0.060
cg20426994	84	0	0.34	0.34	0.031 ± 0.052
cg14692377	77	0.06	0.35	0.29	0.121 ± 0.051
cg21139312	68	0	1	1	0.738 ± 0.244
cg09651136	89	0.01	1	0.99	0.749 ± 0.159
cg20822990	69	0	0.46	0.46	0.220 ± 0.108

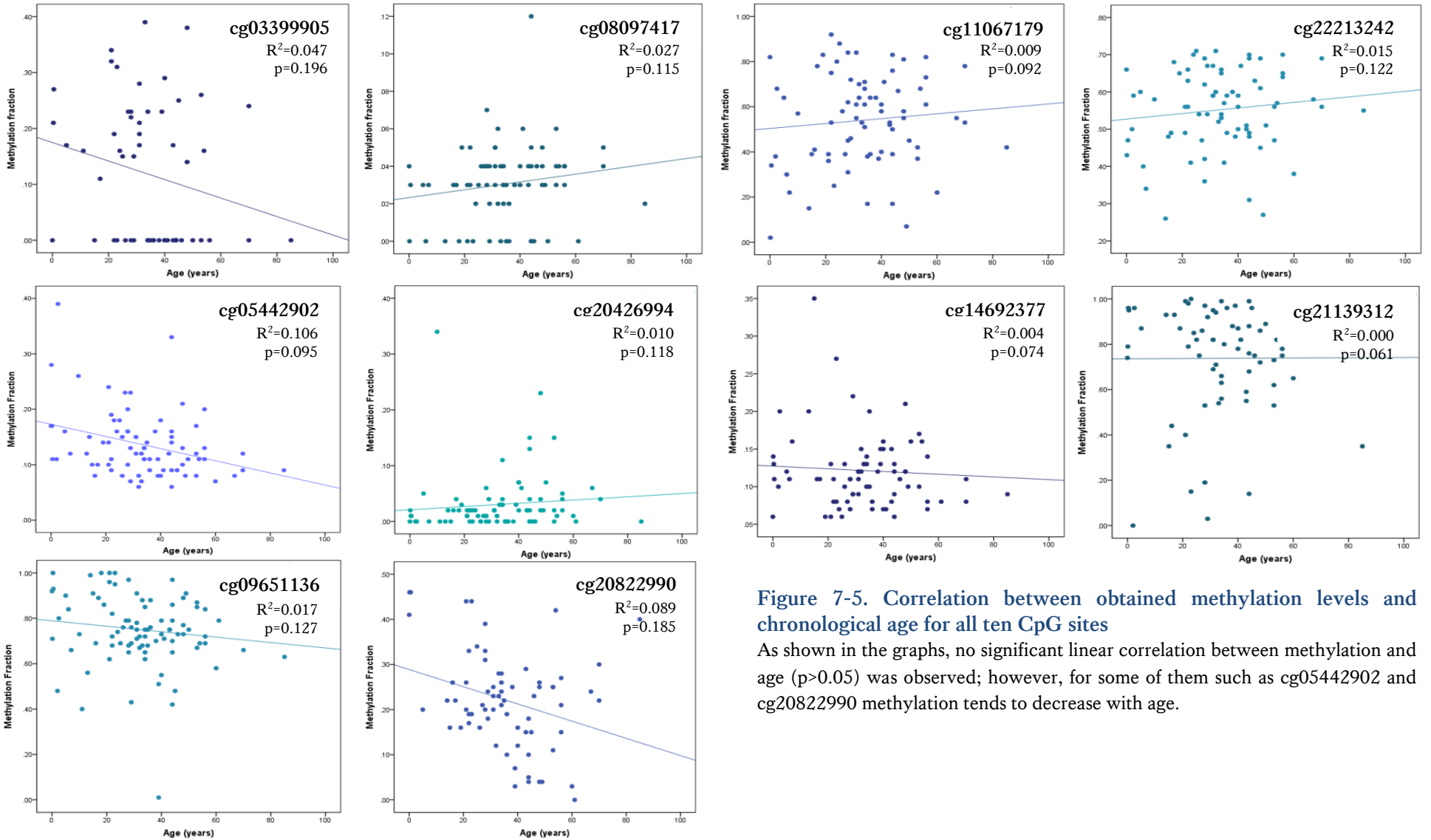


Figure 7-5. Correlation between obtained methylation levels and chronological age for all ten CpG sites

As shown in the graphs, no significant linear correlation between methylation and age ($p>0.05$) was observed; however, for some of them such as cg05442902 and cg20822990 methylation tends to decrease with age.

7.3.1.3 Age prediction

Although the obtained methylation values showed no significant linear correlation with age when taken individually, multiple regression analysis was explored as it could assess all markers' methylation changes together. Occasionally, there were one or two missing methylation values per sample so for this type of analysis a final set of 65 samples with complete data for all markers was used. As shown in Figure 7-6, simple multiple regression involving only a limited number of variables has shown a significant association with chronological age (r=0.54, p=0.025); however, the resulted prediction showed significant error. Two variables, namely cg05442902 and cg20426994, were the most important in predicting age (p=0.002 and p=0.025 respectively).

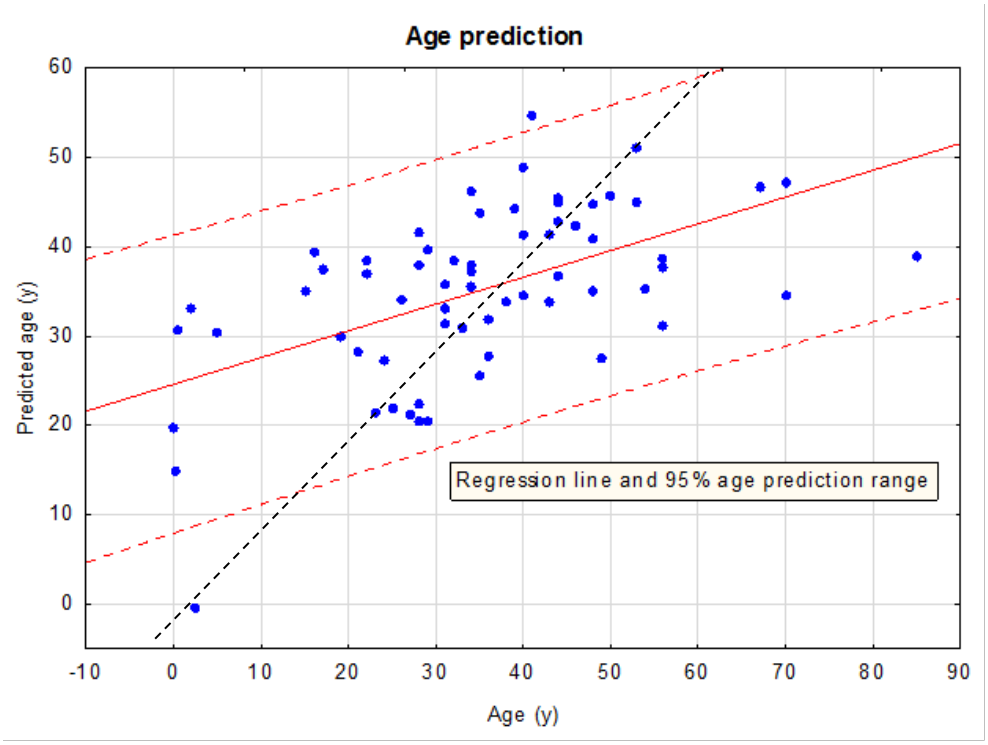
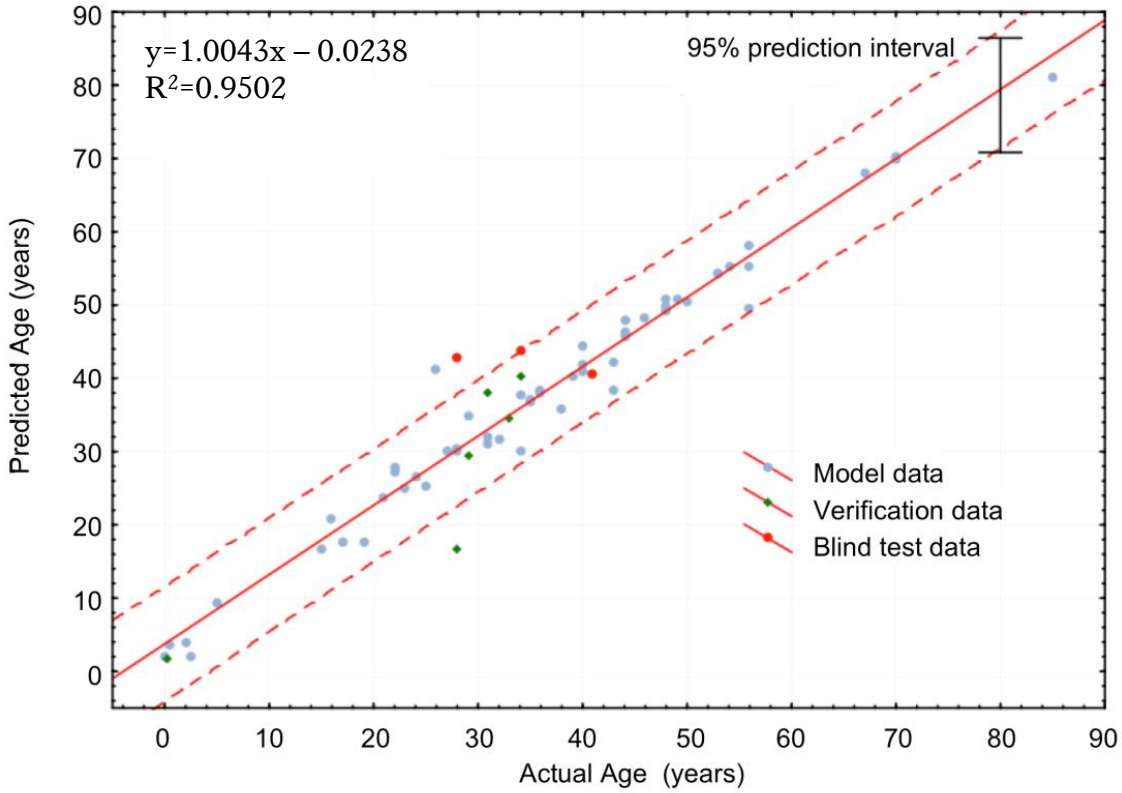


Figure 7-6. Age prediction using multiple regression analysis (n=65, 10 CpGs)
Blue dots symbolise individual age predictions, while the red line represents the regression line with 95% age prediction range (dashed red line). The black dashed line illustrates the 'ideal' prediction. As shown, the resulted regression mostly overestimates young individuals, while underestimates older individuals.

Since some variables were shown to be more important in predicting age than others, it was concluded that the use of neural network analysis could take advantage of the variability in all of these imprecise associations and apply differential weights. Models could be built from a selection of data (training set), which are then verified (verification set) and tested (blind test) from other subsets of available data.

The ageing model, performed by Thomas Miller and presented in Figure 7-7, was trained using the methylation results of a total of 54 blood samples applying artificial neural networks (ANN). The accuracy of the model was high, with a correlation between observed and predicted age of 95% and an average error of 4.27 years. In order to verify the model, the methylation data of six random samples were employed, where the obtained error was 6.09 years. Moreover, when the last three samples were used as a blind test, the error was significantly higher (10.36 years).



ANN development and analysis by T. Miller

Figure 7-7. Age prediction model created by applying ANN (n=65, 10 CpGs)
Methylation data of 65 whole blood samples of individuals aged 1 week to 85 years were employed to create an age prediction model using ANN (mean absolute error=4.27 years in the training set). Different-coloured dots correspond to the training (blue), verification (green) and blind test (red) data. The red dashed lines indicate the 95% prediction interval.

Although the obtained results were promising indicating the potential usefulness of the proposed age-associated DNA methylation markers, the limitations of the method including the small data set could not be overlooked. Furthermore, since issues with amplification bias during bisulphite PCR have been previously observed and the observed methylation variation was in some cases very small, it was decided that a linearity analysis for the proposed Pyrosequencing® assays should be performed before any conclusions on the age prediction model can be made.

7.3.1.4 Linearity of methylation quantification by Pyrosequencing®

As shown in Chapter 5, non-linear methylation quantification is common in bisulphite PCR-based protocols; since there are two different DNA templates available (methylated/unmethylated), PCR efficiencies could differ resulting in amplification bias. To assess the performance of the designed assays regarding methylation quantification, DNA methylation controls (0-100%) (EpigenDx) were used. 100 ng of each standard were bisulphite-treated in duplicate and amplified using all assays (AGE1-9); the average methylation per standard was then calculated [Table 7-8].

Table 7-8. Mean observed methylation values of the tested DNA controls.
The grey box represents missing data.

CpG sites	DNA methylation standards						
	0	0.05	0.1	0.25	0.5	0.75	1
cg03399905	0.01	0.03	0.09	0.19	0.39	0.61	0.96
cg08097417	0.01	0.02	0.10	0.34	0.45	0.65	0.80
cg11067179	0.00	0.13	0.26	0.45	0.59	0.72	0.81
cg22213242	0.00	0.10	0.32	0.48	0.49	0.68	
cg05442902	0.06	0.12	0.20	0.28	0.46	0.57	0.58
cg20426994	0.01	0.15	0.12	0.32	0.49	0.62	0.90
cg14692377	0.01	0.04	0.10	0.18	0.32	0.28	0.46
cg21139312	0.06	0.20	0.18	0.25	0.46	0.60	0.95
cg09651136	0.04	0.08	0.13	0.37	0.72	0.80	0.99
cg20822990	0.02	0.12	0.33	0.53	0.73	0.81	0.86
Average	**0.02**	**0.10**	**0.18**	**0.34**	**0.51**	**0.63**	**0.81**

While the unmethylated DNA (<0.05) was correctly sequenced in all assays (mean detected methylation=0.02), issues regarding the methylated standard (>0.85) were observed for certain CpGs. For example, for cg05442902 and cg14692377, the average methylation values were 0.58 and 0.46 respectively. Interestingly, the maximum obtained methylation values for these particular CpGs when analysing the blood samples were also 0.39 and 0.35 [Table 7-7]. Furthermore, as a general statement, it was noted that there was bias towards the methylated allele in low methylated standards (5-25%), while the opposite bias towards the unmethylated allele was observed in highly methylated standards (75-100%). This phenomenon resulted in methylation over-estimation and underestimation respectively. Taking into account the above, a graph of observed *vs.* 'normalised' expected methylation for each marker was drawn and the best-fitted trendline (quadratic or cubic polynomial) was chosen as previously discussed in section 2.4.1.5 [Figure 7-8].

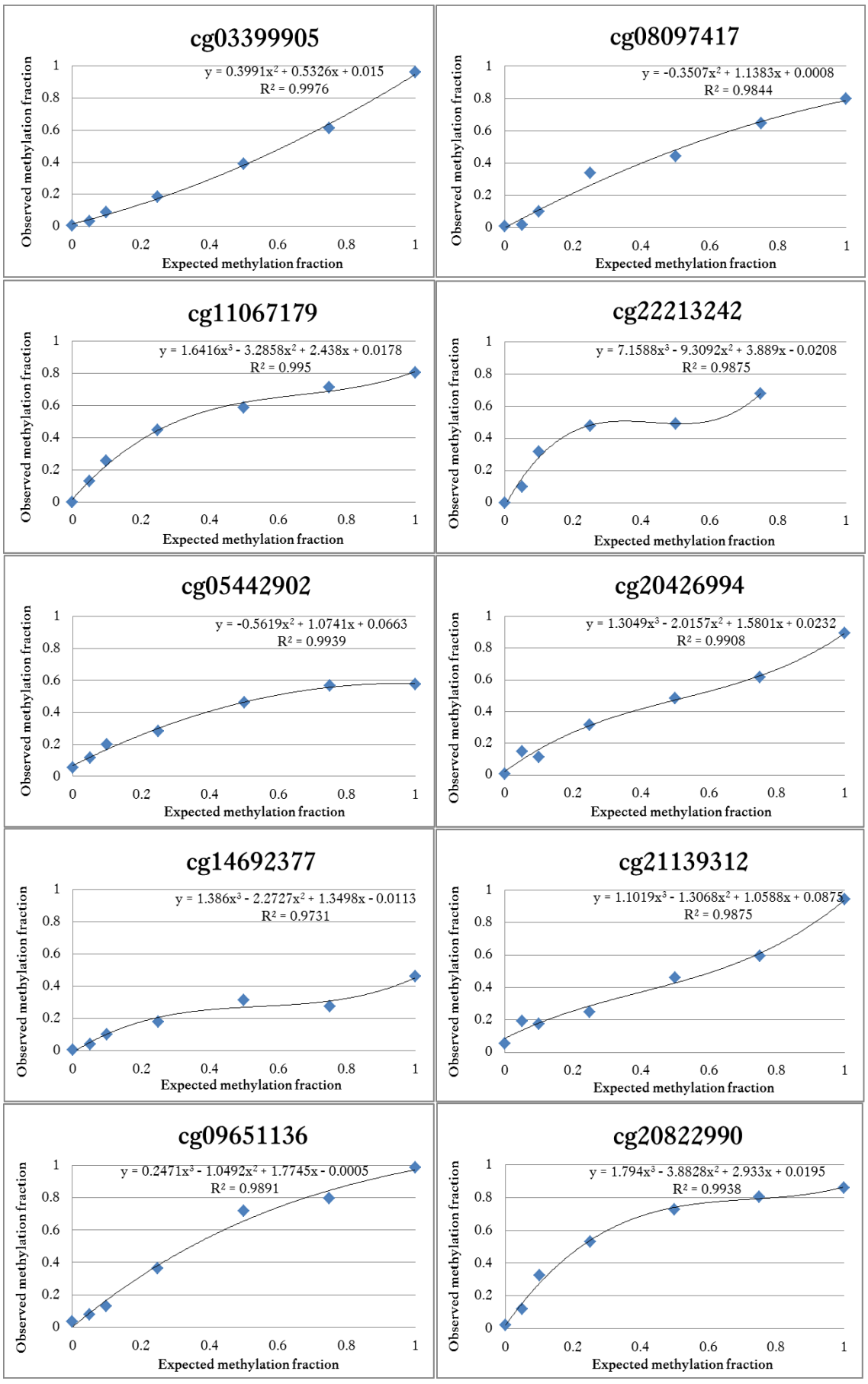


Figure 7-8. Observed vs. expected methylation of known DNA methylation standards for all ten CpG sites

7.3.1.5 Verification of the developed model

In an attempt to verify these results, the same blood samples were reanalysed using the same experimental conditions in order to fill any 'gaps' in the previous methylation dataset; the final dataset consisted of 86 blood samples with methylation data for all ten CpG sites in question. Using the equations derived from the linearity analysis, the methylation values of these blood samples were also 'corrected' as described in section 2.4.1.3. Both the original and corrected methylation datasets were employed and reanalysed to assess if there was an improvement in age prediction.

Again, simple multiple regression analysis did not result in a significant correlation with age ($r=0.44$, $p=0.08$ using the original and $r=0.39$, $p=0.19$ using the corrected methylation values); however, using analysis of variance (ANOVA) the importance of each marker could be considered individually. Using a stepwise model it was confirmed that cg20822990 was highly associated with age in both datasets ($p=0.002$ for the original and $p=0.023$ for corrected data) followed by cg05442902 ($p=0.039$ for the original and 0.048 for the corrected data); the other age variables did not seem to improve the model. These two markers alone could explain 15% of variation in age in this sample set. Interestingly, cg05442902 seemed to demonstrate the highest correlation with age also in the previous multiple regression model [Figure 7-6]. These results were also verified by univariate analysis where age prediction using all ten CpG sites was not drastically better than when using only the cg05442902 and cg20822990 for the uncorrected data [Figure 7-9]. However, the age prediction resulted in a high standard error of estimate of 16-17 years in all models tested. Surprisingly, similar results were also obtained using neural networks where creating an accurate prediction model was rather challenging.

In general it was observed that the correction of the methylation quantification using the equations obtained by analysing known DNA methylation standards did not improve age prediction; the correlation was actually better using the original methylation values obtained by the pyrosequencer. It is believed that this is due to the low range in methylation levels observed for certain markers; therefore, correction did not actually change the obtained values. Also, the resolution of the method itself should be taken into account since there was an average difference of detected methylation between replicates of 6.5%.

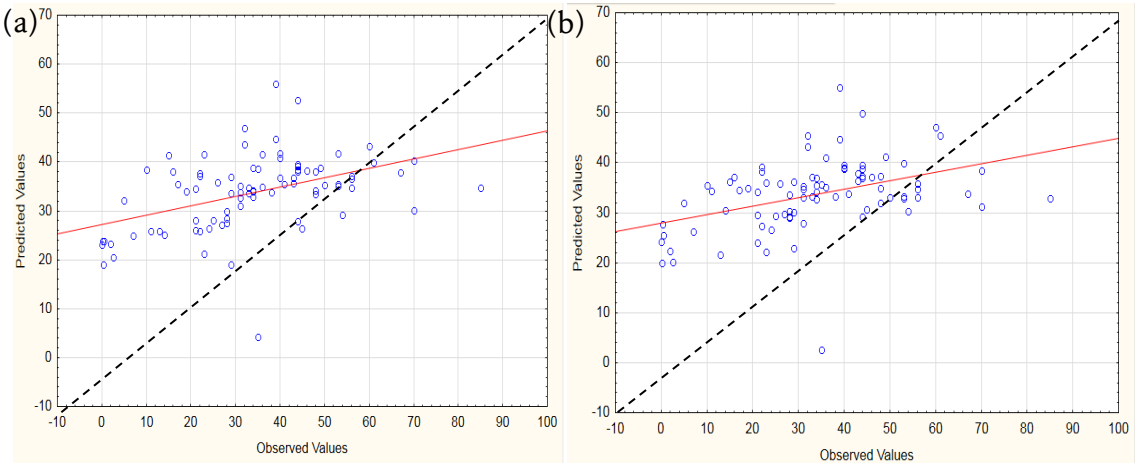


Figure 7-9. Age prediction by univariate analysis in the original methylation data (n=86)
(a) Age prediction using all ten predictors, (b) Age prediction using only cg05442902 and cg20822990. Red lines indicate linear regression while the dashed black lines show the expected age prediction.

Failing to reproduce the age prediction results using a slightly larger dataset could not only be due to experimental parameters and the method's accuracy, but could also indicate that applying only the ten selected age-associated CpG sites is not enough to predict age. It is known that ANN models predict better using fewer cases, hence the promising age prediction after the first attempt [Figure 7-7]. It is believed that a larger set of CpG sites gathered by a larger genome-wide methylation analysis containing thousands of samples could potentially overcome existing challenges and reveal more robust age-associated markers.

7.3.2 Age prediction model through artificial neural networks (ANN)

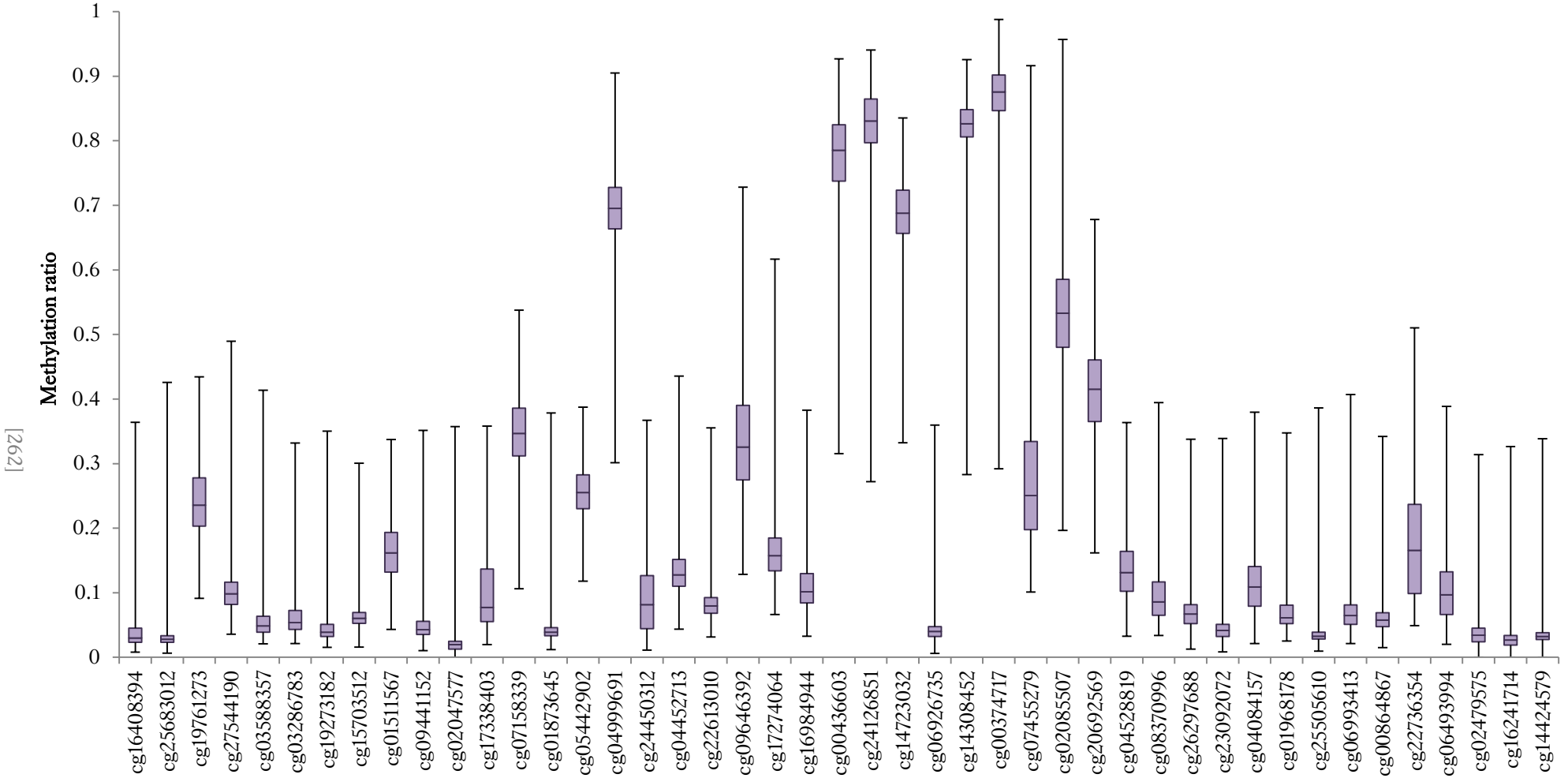
7.3.2.1 Age-associated DNA methylation changes in blood

In an attempt to test more robust CpG sites, a set of 45 CpG sites were selected from the study by Horvath (2013) according to the criteria already mentioned in section 7.2.2.2 [Table 7-6]. Methylation data regarding these CpG sites were collected from a total of 1,156 whole blood samples from individuals 2-90 years old included in seven, publicly available methylation databases. Box and whisker plots were obtained for each essay that indicate the median, quartiles (1st and 3rd) as well as methylation range for each assay [Figure 7-10]. As expected, some CpG sites showed higher variation than others; cg07455279 demonstrated the highest methylation range (0.815) while cg05442902 usually showed low methylation levels (0.118 to 0.387). Since the latter is the only common marker with the previous approach, it is worth noting that the methylation pattern matches with that previously reported via Pyrosequencing® (0.06-0.39) [Table 7-9].

Methylation fractions (zero to one) were then plotted against the actual age of each individual in order to investigate potential correlation between the methylation levels and age. In general, the methylation of certain CpG sites such as cg19761273, cg01511567, cg07158339 and cg05442902 was clearly decreasing with advancing age, while some others, cg20692569, cg04528819, cg04084157 and cg22736354 to name but a few were increasingly methylated over time. These observations align with the age relationship that Horvath reported in his study. Interestingly, at least 16 CpG sites (e.g. cg02047577 and cg01873645) were found constantly unmethylated (<0.2), while three CpG sites – cg24126851, cg14308452 and cg00374717 – demonstrated high methylation levels for all ages.

7.3.2.2 Identification of the epigenetic ageing signature

The observed age-associated methylation changes were assessed for their statistical significance for all markers in an attempt to identify the ones that could form the proposed epigenetic-ageing-signature. Firstly, using multivariate analysis and testing the effect of gender or ethnicity on age-associated methylation, no significant correlation was determined (p=0.77102 and p=0.091548 respectively).



[262]

Figure 7-10. Observed methylation variation for all 45 CpG sites

Using linear regression analysis, a significant correlation between methylation levels and age ($p<0.05$) was confirmed for 25 out of the 45 CpG sites (shaded in green on Table 7-9). As expected, most of these markers are the ones that were identified earlier for having a decreasing or increasing trend in their methylation status with advancing age; however, using multiple regression analysis, it was observed that not all markers significantly improved age prediction. Applying stepwise regression for variable selection, the results regarding the importance of the markers was similar, although the order of the markers was slightly different [Table 7-10]. In order to perform this type of analysis, the markers were added one by one into the age prediction model until there was no statistical improvement. As shown in Table 7-10, the addition of 23 CpG sites resulted in a value of $R^2=0.923$, which was not further improved with the addition of more markers. All 23 age-associated CpG sites that were revealed following stepwise regression are also included in the set of markers identified after multiple regression analysis. Interestingly, in both analyses the marker cg22736354 was found the strongest one and the marker that contributes to age prediction the most.

Table 7-9. Correlation of selected CpG sites with age as assessed by their p-values

CpG sites	p value
cg22736354	0.0000001
cg06493994	0.0000001
cg19761273	0.0000001
cg04528819	0.0000001
cg04084157	0.0000001
cg20692569	0.0000001
cg02085507	0.0000001
cg01511567	0.000002
cg27544190	0.000006
cg05442902	0.000044
cg17274064	0.000186
cg14308452	0.000265
cg16408394	0.000601
cg07158339	0.001582
cg02479575	0.002700
cg23092072	0.010637
cg08370996	0.018122
cg04999691	0.02183
cg24126851	0.026452
cg22613010	0.027471
cg17338403	0.029662
cg04452713	0.034067
cg19273182	0.03497
cg24450312	0.042915
cg03286783	0.049725
cg06993413	0.180880
cg14424579	0.198533
cg14723032	0.218067
cg06926735	0.227325
cg03588357	0.305084
cg09646392	0.312717
cg07455279	0.424543
cg01968178	0.427676
cg16984944	0.436248
cg16241714	0.453502
cg25505610	0.488388
cg02047577	0.509941
cg01873645	0.608939
cg26297688	0.629988
cg00864867	0.718095
cg15703512	0.723257
cg25683012	0.813015
cg09441152	0.964471
cg00374717	0.988213
cg00436603	0.995672

Table 7-10. Summary of stepwise regression for the first 28 CpG sites used in the model
The markers contributing to the model are highlighted in green, while the ones in red are those that do not significantly change age prediction.

CpG sites	Step +in/-out	Multiple R	Multiple R^2	R^2 change	p value
cg22736354	1	0.835200	0.697559	0.697559	0.0000001
cg19761273	2	0.910351	0.828740	0.131181	0.0000001
cg20692569	3	0.920375	0.847090	0.018350	0.0000001
cg06493994	4	0.928244	0.861636	0.014547	0.0000001
cg27544190	5	0.939846	0.883311	0.021674	0.0000001
cg17274064	6	0.943643	0.890461	0.007151	0.0000001
cg04084157	7	0.946874	0.896571	0.006110	0.0000001
cg04528819	8	0.949625	0.901787	0.005216	0.0000001
cg01511567	9	0.951676	0.905687	0.003900	0.0000001
cg02085507	10	0.953522	0.909204	0.003517	0.0000001
cg07158339	11	0.954730	0.911510	0.002306	0.000001
cg05442902	12	0.956017	0.913968	0.002458	0.0000001
cg02479575	13	0.956845	0.915552	0.001583	0.000030
cg08370996	14	0.957458	0.946727	0.001175	0.000290
cg24450312	15	0.957983	0.917732	0.001005	0.000744
cg03286783	16	0.958845	0.919383	0.001652	0.000013
cg23092072	17	0.959382	0.920414	0.001030	0.000526
cg14308452	18	0.959720	0.921063	0.000650	0.005678
cg16408394	19	0.959903	0.921414	0.000350	0.041706
cg24126851	20	0.960074	0.921742	0.000329	0.048114
cg04452713	21	0.960309	0.922194	0.000452	0.020294
cg22613010	22	0.960498	0.922556	0.000362	0.037302
cg17338403	23	0.960724	0.922990	0.000434	0.022426
cg19273182	24	0.960791	0.923119	0.000129	0.212676
cg04999691	25	0.960858	0.923247	0.000129	0.213073
cg06993413	26	0.960937	0.923400	0.000153	0.170420
cg06926735	27	0.961028	0.923575	0.000174	0.146738
cg14723032	28	0.961108	0.923728	0.000154	0.172929

The methylation change over time regarding the first 16 CpG sites proposed by the stepwise regression analysis is also illustrated in Figure 7-11. Information regarding the genes that the CpG sites lay near or within was acquired to identify their function and potential involvement in ageing. The exact chromosomal locations of the CpG sites comprising the epigenetic ageing signature as well as the involved genes are also shown in Table 7-11.

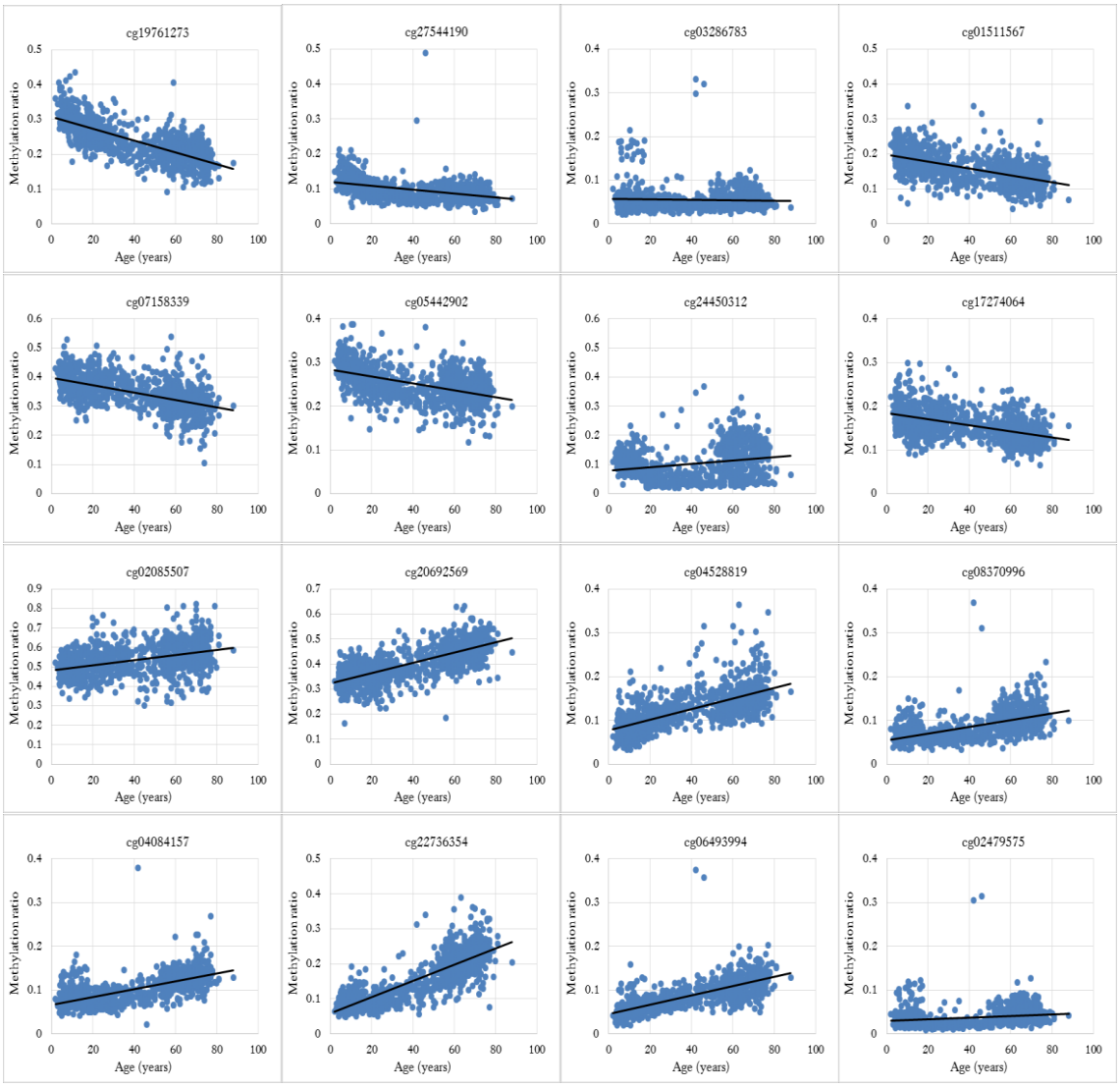


Figure 7-11. Change of methylation levels over advancing age for the 16 most important age-associated CpGs (n=1,156)

Each dot represents the observed methylation of an individual for the CpG site in question. It should be noted that in these graphs age is shown as age-in-months. Black lines correspond to linear regression.

7.3.2.3 Age prediction in blood

Applying multiple regression analysis on the methylation values of all 1,156 individuals for the selected 16 best CpG sites using the STATISTICA software, the correlation between predicted and true age was significantly strong (linear correlation, R^2=0.9258), while the mean absolute error using all data was 4.89 years (standard deviation = 4.36 years) [Figure 7-12]. As shown in the graph, 60.5% (700/1,156) of individuals were predicted within a ±5 year error range, while 89% (1,029/1,156) samples were predicted within a ±10 year error range. As indicated in Figure 7-12, there were individuals that seemed to age fast (red circles) while there were others that were predicted as much younger (green circles). Notably, as shown in Figure 7-12b,

the prediction error in older individuals (>60 years old) was higher compared to younger ones. This is believed to be expected as older individuals have been exposed to more environmental stress throughout their lifetime that could have potentially caused changes in DNA methylation patterns (epigenetic drift).

Table 7-11. Epigenetic ageing signature consisting of 16 CpG sites

CpG sites	Chromosomal location	Gene
cg19761273	17: 80,232,096	CSNK1D - casein kinase 1; delta isoform 1
cg27544190	21: 33,785,434	C21orf63 - chromosome 21 open reading frame 63
cg03286783	15: 44,580,973	CASC4 - cancer susceptibility candidate 4 isoform a
cg01511567	11: 57,103,631	SSRP1 - structure specific recognition protein 1
cg07158339	9: 71,650,237	FXN –frataxin, mitochondrial isoform 1 preproprotein
cg05442902	22: 21,369,010	P2RXL1 - purinergic receptor P2X-like 1; orphan receptor
cg24450312	1: 206,681,158	RASSF5 - Ras association domain family 5 isoform B
cg17274064	21: 40,033,892	ERG - v-etserythroblastosis virus E26 oncogene like isoform 2
cg02085507	19: 6,739,192	TRIP10 - thyroid hormone receptor interactor 10
cg20692569	7: 72,848,481	FZD9 - frizzled 9
cg04528819	7: 130,418,315	KLF14 - Kruppel-like factor 14
cg08370996	15: 96,874,031	NR2F2 - nuclear receptor subfamily 2; group F; member 2
cg04084157	7: 100,809,049	VGF - nerve growth factor inducible precursor
cg22736354	6: 18,122,719	NHLRC1 – malin
cg06493994	6: 25,652,602	SCGN - secretagogin precursor
cg02479575	19: 4,769,653	C19orf30 - hypothetical protein LOC284424

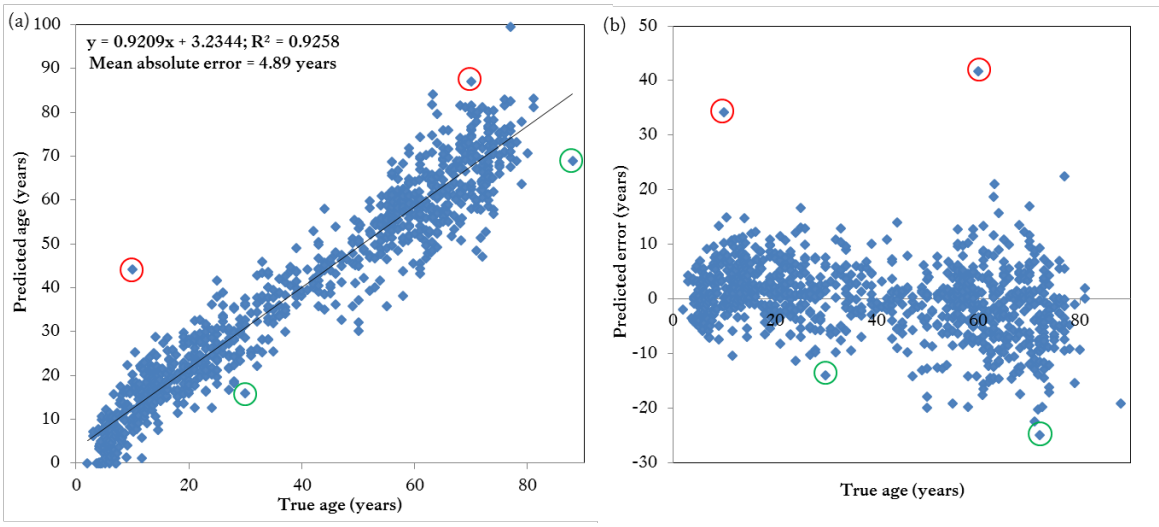
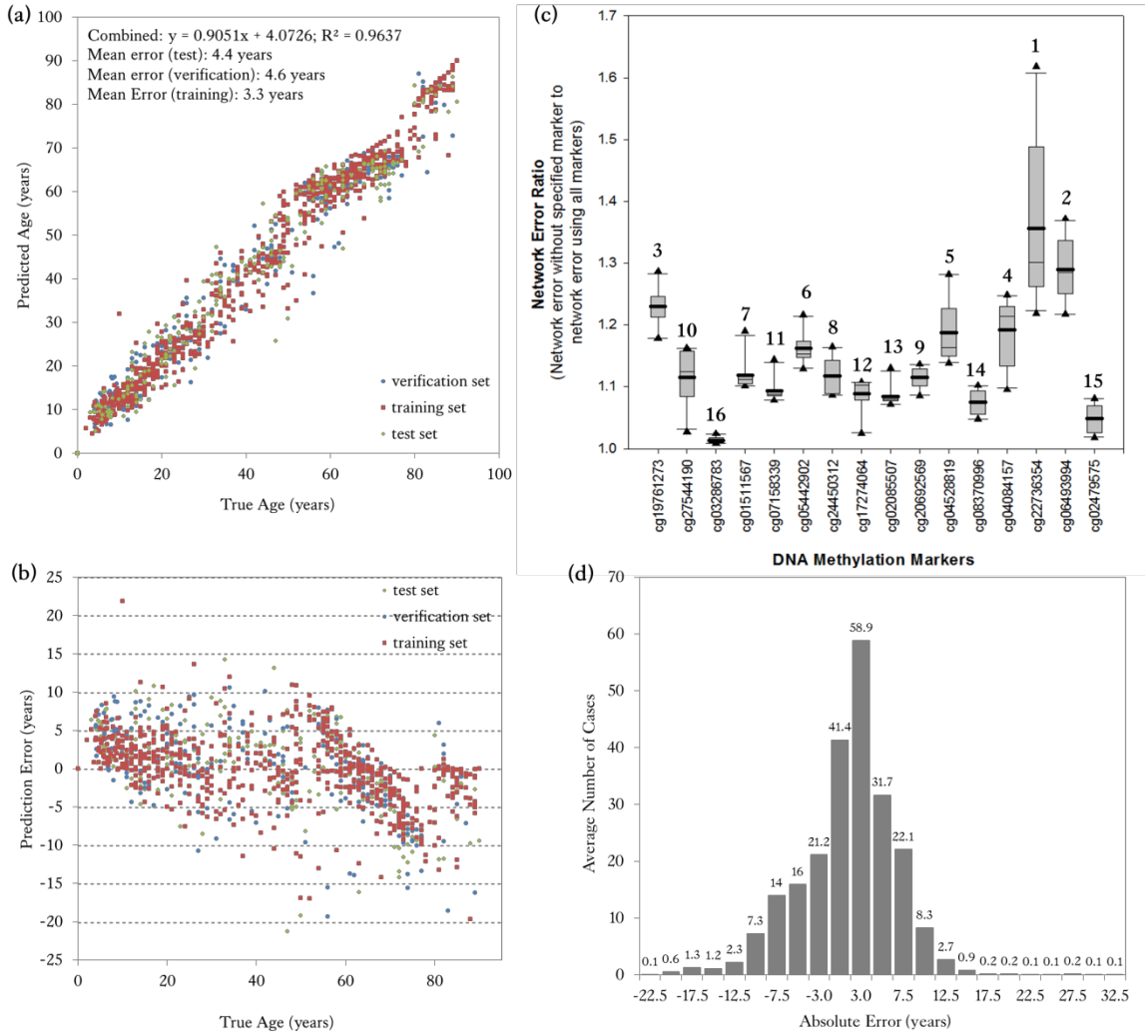


Figure 7-12. Age prediction using multiple regression analysis (16 CpG sites)
(a) Predicted *vs.* Chronological age (years) for all 1,156 individuals used in this study (linear correlation R^2=0.92, mean absolute error=4.89 years, standard deviation=4.36 years), (b) Predicted error (years) over advancing age. As shown most individuals were predicted within a ±5 year error range (0.61), while 1,029 out of 1,156 samples were predicted within a ±10 year error range (0.89). Red circles indicate individuals that were anticipated to age fast, while green circles indicate individuals that were predicted as being much younger.

The potential of applying artificial neural networks was once again explored since it is believed that it could result in a better prediction, especially for complex variables like age. For the purpose of this study, two different ANN models were investigated including a multilayer perceptron (MLP) and a generalised regression neural network (GRNN). Applying the first approach using the standard settings specified within the STATISTICA software, the age prediction was slightly better (mean absolute error=3.54 years, standard deviation=2.9 years, training) than the one obtained by regression analysis (mean absolute error=4.9 years, standard deviation=4.4 years).

On the other hand, neural network models developed by Dr Leon Barron using Trajan 6.0 software showed that the age prediction accuracy could be significantly improved using these data. This prediction was also shown to be repeatable and is presented as an average of the predicted values acquired from n=10 replicate networks from the optimised GRNN architecture. The results were promising as they resulted in a correlation between predicted and true age of 96% and a mean absolute error of 3.3 years (standard deviation=3.7 years) [Figure 7-13a]. In more detail, a total of 694 samples were used to train the model, while 231 samples were used for each of the verification and blind tests (for a total of 1,156 samples). Predictions were still highly accurate for the both test sets (error of 4.6 and 4.4 years correspondingly). As shown in Figure 7-13b, 95.6% of the samples was predicted within 10 years, while 82.8% were predicted within 5 years. Interestingly, the graph showing the age residual error obtained by ANN (Figure 7-13b) has a noticeably different pattern than the one obtained by multiple regression analysis (Figure 7-12b). While the latter resulted in a more randomly distributed error pattern around the mean, errors obtained by ANN revealed at least two (or potentially three) different patterns within the graph. Looking at the graph (Figure 7-13b], one can notice that there are parts within the graph showing that predicted errors decrease with age (even demonstrating a linear trend). These findings could indicate a potential ANN bias in age prediction within specific age groups that could perhaps be corrected. Furthermore, Figure 7-13c shows a sensitivity analysis on the contribution of each marker to the age prediction produced by the GRNN models, where the network error ratio upon systematically removing one CpG at the time was assessed. Interestingly, the order that ANN categorised the CpG sites resembled the one obtained by multiple regression and stepwise analysis [Tables 7-9 and 7-10].



ANN development and analysis by L. Barron

Figure 7-13. Age prediction using artificial neural networks (16 CpG sites)

(a) Predicted vs. Chronological age for all 1,156 individuals included in the study as calculated by averaging the prediction values of 10 replicate networks (linear correlation R^2=0.96, mean error (test) =4.4 years), (b) Predicted error (years) over advancing age; as shown most individuals were predicted within a ±5 year error range (0.83), while 95.6% of the samples were predicted within a ±10 year error range, (c) This graph represents the network error ratio upon systematic removal of each CpG site from 10 replicate networks. Boxes include data from the 25th-75th percentile as well as the median (thin line) and mean (thick line); error bars include the 5th and 95th percentile. As shown the most useful age-associated CpG site was found to be cg22736354, while the least useful was cg03286783, (d) The graph represents the absolute error in years for the blind test only.

7.3.2.4 Validation through an independent cohort of monozygotic twins

The obtained model was validated using a secondary independent cohort, consisting of an additional 106 blood samples of 53 monozygotic twin pairs aged 33-77 years. Monozygotic twins were chosen since they begin life with nearly identical genetic and epigenetic profiles and it is the effect of various environmental factors that alters their genome-wide DNA methylation profile later in life. Therefore, it would be possible to

assess these types of variations and conclude if age acceleration is highly heritable. The methylation values of each sample for all 16 CpG sites were imported into the model as a blind test (the identity of the twin pairs was not indicated) and the obtained overall prediction errors are shown in Table 7-12.

In general, the obtained age prediction accuracy was lower than the one obtained from the blind test with an average mean absolute error of 7.07 years (standard deviation=5.78 years). More specifically, regarding the 20 twin pairs analysed on the Illumina 450K platform the mean error was 9.86±7.16 years, while for the 23 twin pairs analysed on the Illumina 27K platform the mean error was 5.38±3.97 years. Interestingly, there were twin pairs that were predicted to be either much older or much younger than their actual age, but the differences within the twin pairs were not statistically significant as obtained by paired t-test analysis (p=0.99). The mean difference of age prediction between twins was 2.65±2.37 years. These results are very interesting since they indicate a possible 'inherited' function of ageing and could confirm the belief of 'good genes'. The DNA damage theory of aging proposes that aging is a consequence of the accumulation of naturally occurring DNA damages, shown as DNA alterations, which also include the mechanism of DNA methylation. Since it is accepted that DNA methylation patterns adjust according to environmental exposure, this could potentially indicate that the response mechanisms involved have a genetic basis. For example, the DNA repair processes are regulated by DNA repair enzymes that can potentially have different efficiencies depending on the isoform and the inherited DNA sequence in question. Furthermore, one can assume that twins within the same pair are more likely to be exposed to similar environmental conditions than twins from different pairs, therefore the differences in DNA methylation (as translated in age prediction) within the same pair is expected to be smaller than the differences amongst different twin pairs.

Also, considering that the twins (apart from one pair) were all >45 years old, the effect of environmental conditions and lifestyle should also be considered. According to Horvath, while the heritability of age acceleration was found to be 100% in new-borns, it was only 39% in older subjects suggesting that non-genetic factors become more relevant later in life (Horvath, 2013). Also, although all twins were volunteered as healthy controls, it would be very beneficial if information regarding disease

susceptibility was available. It is believed that together with differences in lifestyle, it could possibly explain the observed large differences in age prediction between twins in four pairs highlighted in black in Table 7-12.

Table 7-12. Age prediction in an independent cohort of monozygotic twin pairs

Twin pairs (1-53) are presented below in age order (33-77 years old). 'X' represents the age prediction error of each twin, which can be classified in five categories (0-2.5, 2.5-5, 5.-7.5, 7.5-10 and>10 years). Results in orange correspond to 20 pairs (Illumina 450K platform, mean error of 9.86±7.16 years), while results in blue correspond to 23 pairs (Illumina 27K platform, mean error of 5.38±3.97 years).

Twin pairs	Age (years)	0-2.5	2.5-5	5-7.5	7.5-10	>10
1	32.95					XX
2	45.59					XX
3	45.85	XX				
4	47.28	X	X			
5	48.3	X				X
6	49.08				X	X
7	49.14					XX
8	49.2		X	X		
9	49.6				X	X
10	50.11		X			X
11	50.82				X	X
12	50.98			X	X	
13	51.55			X	X	
14	51.58				XX	
15	52.96	X		X		
16	54.2		X	X		
17	54.31	X	X			
18	54.55				X	X
19	54.71			X		X
20	54.86		X		X	
21	55.29			XX		
22	55.36			XX		
23	56.41			XX		
24	56.71	XX				
25	56.77			X		X
26	57.33	XX				
27	58.31	XX				
28	58.49		X	X		
29	58.59	XX				
30	59.08	X				X
31	59.28					XX
32	59.6		XX			
33	59.9	XX				
34	60.06			X	X	
35	61.91			X		X
36	62.75					XX
37	63.61		X		X	
38	63.84	X	X			
39	63.9	XX				
40	64.35	X	X			
41	64.38	X	X			
42	64.85	XX				
43	65.47					XX
44	65.52					XX
45	65.55	X	X			
46	66.79					XX
47	67.08	X			X	
48	67.7	X	X			
49	68.13	X	X			
50	68.26		XX			
51	69.3			X	X	
52	75.1					XX
53	76.5	X		X		
Number of predictions		29	18	18	13	28

7.3.2.5 Applying the age prediction model in diseased tissues

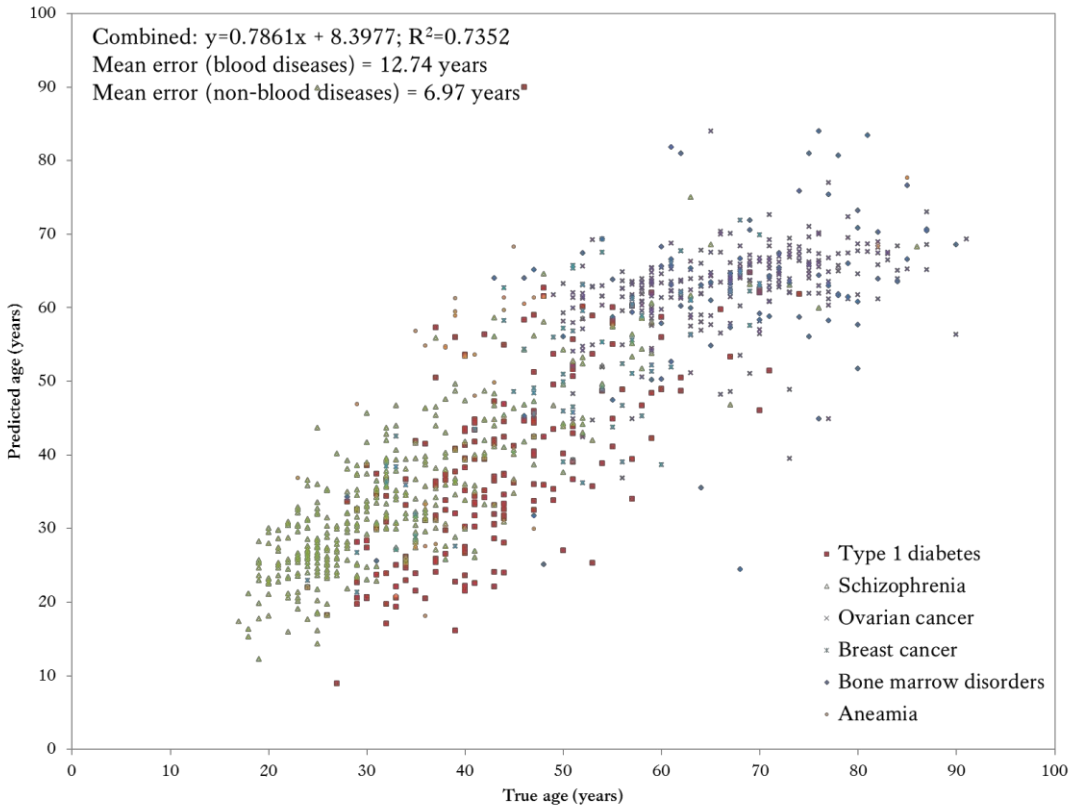
Although Horvath reported that the predicted age from cancer tissues correlated poorly with patient age in his study, six datasets including diseased samples were analysed in an attempt to further validate the proposed age prediction model [Table 7-5]. As previously mentioned, one has to bear in mind that in contrast with a medical setting, information regarding possible disease status is usually not available when trying to predict chronological age from an unknown blood stain during a criminal investigation. Consequently, it is important to build a robust age prediction model containing DNA methylation markers that would not show differential methylation patterns due to diseases. Figure 7-14 shows the predicted *vs.* chronological age for all 1,011 samples included in this analysis; combining all diseases together, a correlation of 73.5% and a mean absolute error of 7.18 years was obtained. However, when analysing separately samples suffering from blood and non-blood diseases it becomes evident that the error is higher for blood diseases (error=12.74 years). This is expected since the methylation data were gathered by analysing whole blood samples. Also, it was noted that as expected (due to the higher environmental stress), the prediction error for older individuals (>60 years old) was higher compared to the younger ones.

In more detail, the obtained mean absolute errors for each disease were as follows:

- **Type 1 diabetes** – error=8.63±6.38 years
- **Anaemia – error**=14.38±5.29 years
- **Bone marrow disorders (including leukaemia)** – error=11.09±8.39 years
- **Ovarian cancer** – error=7.45±5.79 years
- **Breast cancer** – error=6.77±4.94 years
- **Schizophrenia** – error=5.03±4.90 years

Schizophrenia showed the lowest age prediction error, while anaemia demonstrated the lowest correlation with age. While changes in expression of one of the markers included in the model - the VGF neuropeptide linked with cg04084157 - have been detected in the cerebrospinal fluid of patients with different neurological and psychiatric conditions such as schizophrenia (Huang *et al.*, 2006), it did not seem to affect prediction in blood. It should also be noted that schizophrenia patients consisted the largest dataset test; therefore a better prediction error could also be due to the

greater number of samples. On the other hand, the results regarding anaemia (n=28) come as no surprise since anaemia is one of the most common blood disorders which could add extra 'stress' on the body and especially in blood. Interestingly, cg07158339 is located near the frataxin gene (FXN) which has been associated with selectively and non-covalently interacting with ferric ion Fe (III) to assemble the iron-sulfur cluster (Gentry *et al.*, 2013). Consequently, differential methylation patterns due to the disease status in blood cannot be excluded. Another example includes the ERG oncogene associated with cg17274064, which is an erythroblast transformation-specific transcription regulator typically mutated in myeloid leukaemia (Yi *et al.*, 1997). As shown, the dataset comprised by various bone marrow disorders including leukaemia demonstrated the second largest mean error (11.09 years).



ANN development and analysis by L. Barron

Figure 7-14. Age prediction in diseased samples (n=1,011)

7.3.2.6 Applying the age prediction model in other tissues

As mentioned in the Experimental section, a final set of 566 samples for five selected tissues including saliva, buccal cells, skin, cervix and muscle was assembled using methylation data from eight different genome-wide methylation studies [Table 7-4]. Initially, a subset of this dataset was used to import the methylation values into the

already developed prediction model in blood to assess if accurate prediction was possible. This subset consisted of 102 samples - 19 buccal cells, 20 saliva, 15 skin, 20 cervix and 28 muscle samples. In general, prediction was very poor; as an overall remark it can be noted that there was an over prediction in young ages as well as an underestimation of age in older individuals [Figure 7-15]. These results indicate the need to build age prediction models in each forensically relevant tissue separately; however the main challenge is still to obtain a large methylation dataset for analysis.

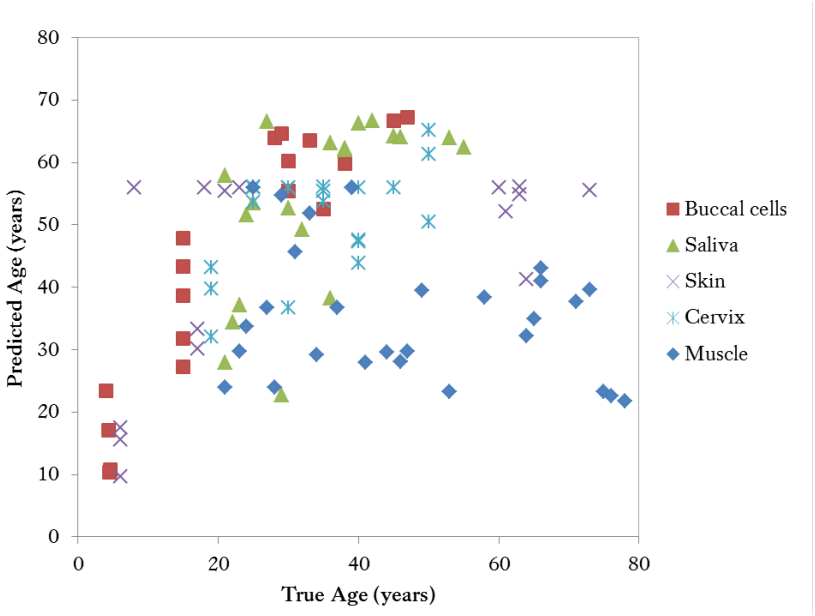


Figure 7-15. Age prediction in other tissues using the developed blood prediction model

7.3.2.7 Age prediction in saliva and cervix

Saliva and cervix were the only tissues where sufficient genome-wide methylation data was available for robust analysis. Issues on building ANN models using less than 100 samples had been previously observed; therefore it was concluded that 19 buccal cells, 53 skin samples and 62 muscle tissues were not enough to build new prediction models. Saliva is one of the most common types of biological evidence found at crime scenes in the form of used glasses, cigarette butts or stamps, whereas vaginal cells are commonly recovered in sexual crime. The purpose of this study was mainly to assess tissue-specific variations in age prediction; therefore, it was decided to focus on these two tissues. For saliva, methylation values regarding the selected 16 CpG sites were collected from a total of 265 samples aged 21-55 years. Similarly, 167 cervix samples from individuals aged 19-69 years were used to build the age prediction for this tissue.

For saliva, 159 samples were used to train a GRNN model, while 53 samples were used for the verification and another 53 for the test set. As shown in Figure 7-16a, a correlation between predicted and true age of 73% was observed while the mean error was 3.18 years (training), 6.26 years (verification) and 4 years (test). The prediction accuracy was great considering the size of the dataset; however, the small age range (21-55 years) cannot be ignored. Also, as shown in the graphs representing the age residuals, age prediction seemed to be more accurate in younger individuals where underestimation of age was not very common, while there was a larger variation in prediction errors for older individuals. On the other hand, the age prediction success in cervix was surprisingly high with a linear correlation of R^2=0.86, while the mean error was 1.01 years (training), 4.44 years (verification) and 5.03 years (test). This should be explained by the fact that, compared to blood or saliva, cervix is a more homogenous tissue made up by distinct cell types. Nevertheless, one should also take into account that the resolution and variation in age was not great for this dataset, which can be easily observed from the distinct lines on the age model [Figure 7-16b].

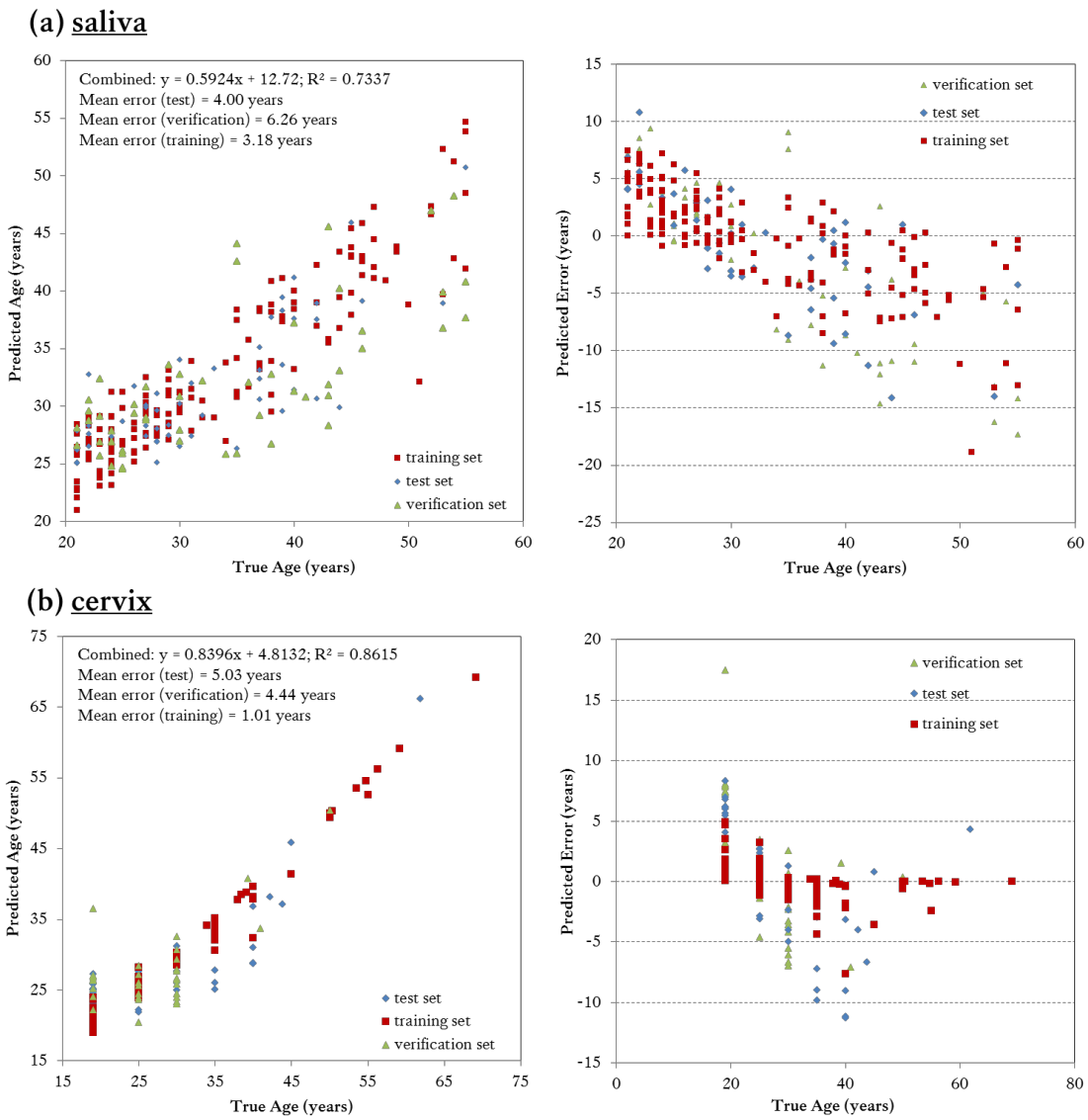


Figure 7-16. Age prediction models in saliva (n=265) and cervix (n=167)

Predicted *vs.* Chronological age (years) for (a) 265 saliva samples (linear correlation R^2=0.73, mean absolute error=4 years (test), standard deviation=3.4 years), (b) 167 cervix samples (linear correlation R^2=0.86, mean absolute error=5.03 years (test), standard deviation=2.97 years)

7.4 Final remarks

Since there have been a few studies that successfully created models using either dozens (Hannum *et al.*, 2013) or hundreds of CpG sites (Horvath, 2013), it was believed that biological age can be predicted using age-associated DNA methylation. From a forensic perspective, the main challenge to be faced is the low quality and quantity of forensic specimens, making it impossible to implement such models in forensic casework in their current form. The purpose of this study was to investigate age-associated CpG sites reported in the literature and assess the possibility of using a small number of markers to accurately predict biological age in blood. To meet the aim, two different approaches were followed.

The first one included the validation of a set of ten CpG sites via bisulphite Pyrosequencing® using 90 blood samples in an attempt to create an age prediction model. Initially, using the methylation data of 65 samples multiple regression analysis revealed that two markers, namely cg05442902 and cg20426994, were the most important in predicting age, however the proposed model lacked accuracy. On the other hand, ANN analysis resulted in an age prediction model with a correlation between observed and predicted age of 95% and an average error of 4.27 years. However, both multiple regression and ANN analysis failed to reproduce the obtained results using a larger dataset. The failure in verifying the genome-wide data could be due to the methodology used. Bisulphite Pyrosequencing® has been effectively applied as a confirmation method in the literature and is considered capable of accurately quantifying DNA methylation levels. However, the resolution of quantification is perhaps not sufficient enough for the purpose of age prediction, where differences in methylation levels are for certain markers very low. Prior to analysis, bisulphite-converted DNA is amplified using 45 PCR cycles which could introduce amplification bias and alter the detected methylation. Using fewer cycles could be a solution; however, this could mean sacrificing the method's sensitivity.

To overcome these challenges, a second approach was designed and a new set of 45 age-associated CpG sites was selected using a recently published age prediction model based on more than 8,000 samples (Horvath, 2013). While the study proposed a total of 353 age-associated CpG sites, only 45 of them were used in this approach. However, in future work all 353 markers could be included, rather than only a subset,

that could potentially reveal more 'strong', suitable markers that could be incorporated into the age prediction model. Since genome-wide methylation data are available via online databases, it was decided that there was no need to re-analyse blood samples using Illumina's platforms in an effort to minimise both cost and time. As a whole, methylation data for these markers from a total of 1,156 blood samples from healthy individuals aged 2-90 years was collected from seven different studies. On a future approach, these variations could be assessed in more detail by investigating available variables such as ethnicity or lifestyle, so that studies can be excluded if they are considered not suitable. Additionally, the way that methylation levels change with advancing age was observed to be quite different for each marker. As Horvath reported, DNA methylation levels only change very gradually with age and the age effects on individual CpGs can be substantially variable.

Simple multiple regression analysis revealed that the methylation levels of 25 out of the 45 markers tested were significantly associated with age ($p<0.05$) and the 16 'best' were used to build a model for age prediction. As expected, all 16 methylation markers are located near or within genes involved in age-related conditions and processes such as DNA repair, cancer and Alzheimer's disease. As shown in Figure 7-12, the correlation between predicted and true age was significantly strong (linear correlation, $R^2=0.9258$), while the mean absolute error using all data was 4.89 years (standard deviation = 4.36 years). As expected, it was observed that subjects age at a different rate with some ageing faster or slower than average, while as expected the higher prediction errors were found in the older age groups (>60 years old).

Age prediction was considerably improved by using artificial neural networks [Figure 7-13]. The results were very promising as they revealed a correlation between age and predicted age of 96% and a mean absolute error of 3.3 years (standard deviation=3.7 years). Predictions were still highly accurate for both the verification and blind test (error of 4.6 and 4.4 years respectively). Overall, 95.6% of the samples were predicted within 10 years, while 82.8% were predicted within 5 years. Considering that similar prediction accuracy was observed when using 353 CpG sites in Horvath's study (age correlation of 0.96 with an median absolute error of 3.6 years), it can be concluded that predicting age using a small number of CpG sites could be possible; however caution is needed when applying in other tissues.

8 Development of a next generation sequencing method for age prediction

8.1 Introduction

The ability to access the epigenetic status of a set of genes or the entire genome deeply facilitates researchers' insight into the nature of gene regulation and various epigenetic mechanisms involved in the interaction between cells and environment. Overall, current methodologies can be divided into genome-wide or gene-specific depending on the number of CpG sites being analysed. As previously mentioned in Chapter 1, they can also be further categorised based on how the methylation status is examined; a) methods that discriminate the bisulphite-stimulated 'cytosine to thymine' transition, b) methods based on DNA cleavage by methylation-sensitive restriction enzymes and lastly, c) methods that employ methyl-binding proteins or antibodies to immunoprecipitate against methylated cytosines (Ammerpohl *et al.*, 2009). Each approach has its own advantages and limitations resulting in either qualitative or quantitative assessment of methylation patterns. Bisulphite-based methods are usually sensitive but have to encounter the reduced genome complexity following DNA conversion that results into technical issues regarding target-specific probe design. On the other hand, methylation-sensitive restriction enzymes cannot often probe every CpG site, whereas immunoprecipitation techniques lack the single-base resolution of a targeted sequence.

Taking into account the sensitivity and accuracy of the available approaches, it was decided that bisulphite-based methods could prove the best choice since they are compatible with technology already available in forensic laboratories. However, most of these methods can be labour-intensive, result in qualitative results and can be limited to the measurement of the methylation status of only one or very few CpG sites. On the other hand Pyrosequencing® technology has proven to overcome some of these limitations requiring only a few nanograms of starting DNA material to obtain high reproducible results (Dejeux *et al.*, 2009). Therefore, since Pyrosequencing® has successfully been used in previous age prediction studies (Weidner *et al.*, 2014) it was chosen thus far to analyse various age-specific differentially methylated markers [Chapter 7].

However, while Pyrosequencing® is useful, it is limited in its quantitative accuracy, read length and sample throughput; the analog nature of its sequence output (light signal) limits its quantitative power. In this study issues regarding reproducibility and precision were observed [Figure 5-7]. In cases where differentiation was based on a DNA methylation marker being in an unmethylated (<0.05) state in one sample while in a highly methylated (>0.85) state in another, one could confidently discriminate them using methylation patterns. Difficulties were raised when the observed methylation differences were much smaller as shown in some of the proposed tissue-specific methylation markers and for most of the age-associated CpG sites [Figure 7-11]. Considering that the obtained bisulphite conversion rates usually varied between 90-100% and that amplification bias in bisulphite PCRs was quite common, it is believed that errors in methylation quantification could be easily introduced. Therefore, even though analysing replicates could serve as a solution, accurate and high-resolution methylation quantification that will allow for discriminating methylation differences of less than 0.05 remains a challenging task.

8.1.1 Next generation sequencing in epigenetics

Searching for a more accurate and higher-resolution sequencing method, next generation sequencing (NGS) technology was considered to be superior. It is generally believed that NGS technology has transformed the way scientists extract genetic information from biological material by introducing sequencing in a massively parallel fashion. NGS allows for rapid sequencing of not only large DNA fragments but even entire genomes in a matter of hours or days. Therefore, technical issues of conventional sequencing such as Sanger sequencing or Pyrosequencing® concerning throughput, scalability, speed, and resolution can be overcome. Also, even though the latest high-throughput sequencing instruments are capable of massive data output, its technology offers high flexibility and scalability but still using the same underlying chemistry. This scalability allows for simultaneous sequencing of hundreds of targeted regions for a large number of samples by applying individual 'barcode' sequences to differentiate between them; an overview of the high-throughput science is illustrated in Figure 8-1.

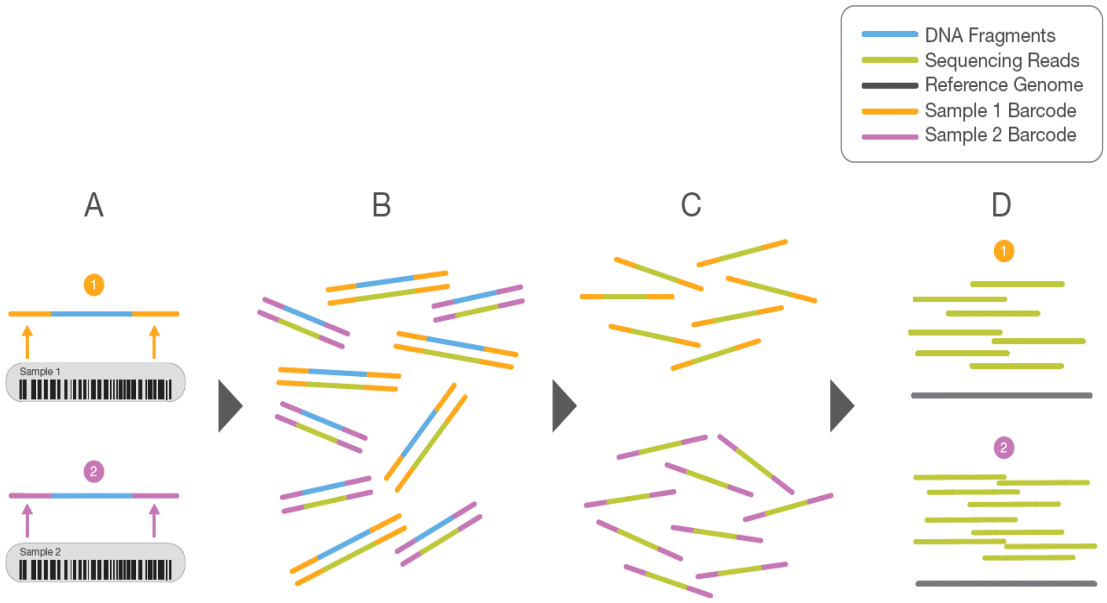


Figure 8-1. Conceptual overview of sample multiplexing (Illumina, 2013a)
(A) Two representative DNA fragments from two different samples, each attached to a specific barcode sequence used to identify the sample from which it originated, (B) Libraries for each sample are pooled together and sequenced in parallel; each new read contains both the fragment sequence and its sample-identifying barcode, (C) Barcode sequences are used to de-multiplex, or differentiate reads from each sample, (D) Each set of reads is aligned to the reference sequence.

Next generation sequencing technology was quickly adopted by the epigenetics community. So far, epigenetic approaches based on various NGS platforms include whole-genome bisulphite sequencing, methylation beadchips, reduced representation bisulphite sequencing and methylated DNA immunoprecipitation sequencing. NGS-based methods could overcome common technical problems of commonly-used microarray analysis (Hurd & Nelson, 2009). For example, since bisulphite-treated DNA is directly sequenced and not interrogated by hybridisation to specific sequences, any experimental bias and cross-hybridisation issues from analysis can be eventually eliminated. Also, NGS offers single-nucleotide resolution; therefore researchers can monitor expression from gene alleles that differ in as little as one nucleotide. Most importantly, the quantification of signal from sequence-based approaches is based on measuring sequence tags rather than relative measure between samples resulting in unlimited, fully quantitative dynamic range of signal. Lastly, only a few nanograms of material are sufficient for NGS, reducing the reliance on amplification of material.

The advantages of having single-base resolution have been illustrated in many studies, where genome-wide DNA methylome analysis via the combination of bisulphite DNA with NGS platforms revealed not only tissue-specific but also age-associated

differentially methylated CpG sites (Almen *et al.*, 2014; Thompson *et al.*, 2013). However, genome-wide sequencing methods do require large amounts of DNA by forensic standards; their capability to use trace DNA samples will therefore be crucial to the application of such methods in forensic epigenetic analysis. While whole bisulfitome amplification meaning genome-wide amplification of bisulphite-modified DNA template has been proposed to overcome this problem (Paliwal *et al.*, 2010), it is believed that it could introduce further bias in the methylation quantification. As shown in previous chapters, these studies can be successfully employed to select suitable markers for validation using other instrumentation but it would not be suitable to analyse thousands of CpG sites when only a few need to be investigated.

8.1.2 Next generation sequencing in forensic genetics

Forensic DNA samples are usually of limited quality/quantity often failing to meet the requirements of simultaneously examining multiple genomic loci. This leads to difficulties in providing sufficient information, therefore limiting their use as legal evidence (Yang *et al.*, 2014). Nowadays, the forensic community worldwide is investigating the value and potential of NGS in improving not only DNA testing but also current workflows. Even though there are plenty of different applications for which NGS can be employed (Illumina, 2013a), targeted sequencing is believed to be more relevant for forensic applications. Using this approach, only a subset of genes or defined regions spanning hundreds of base pairs in the genome are sequenced, which could potentially allow forensic scientists to detect significant genetic variations. So far, scientists have applied the sequencing by synthesis (SBS) technology to sequence large number of PCR amplicons allowing for successful analysis of even the smallest, most compromised and highly-mixed biological evidence samples [Figure 8-2].

Potential forensic genomic applications that NGS can be successfully applied include STR typing (Warshauer *et al.*, 2013), mitochondrial genome sequencing (Parson *et al.*, 2013), Y-chromosome analysis (Xue *et al.*, 2009), differentiation of twins (Weber-Lehmann *et al.*, 2014), microbial forensics (Brenig *et al.*, 2010), species identification (Hajibabaei *et al.*, 2011), ancestry studies and phenotypic inferences (Yang *et al.*, 2014). Even though the implementation of such a technology would still need to go through strict validation criteria for use in forensic casework, it seems very promising.

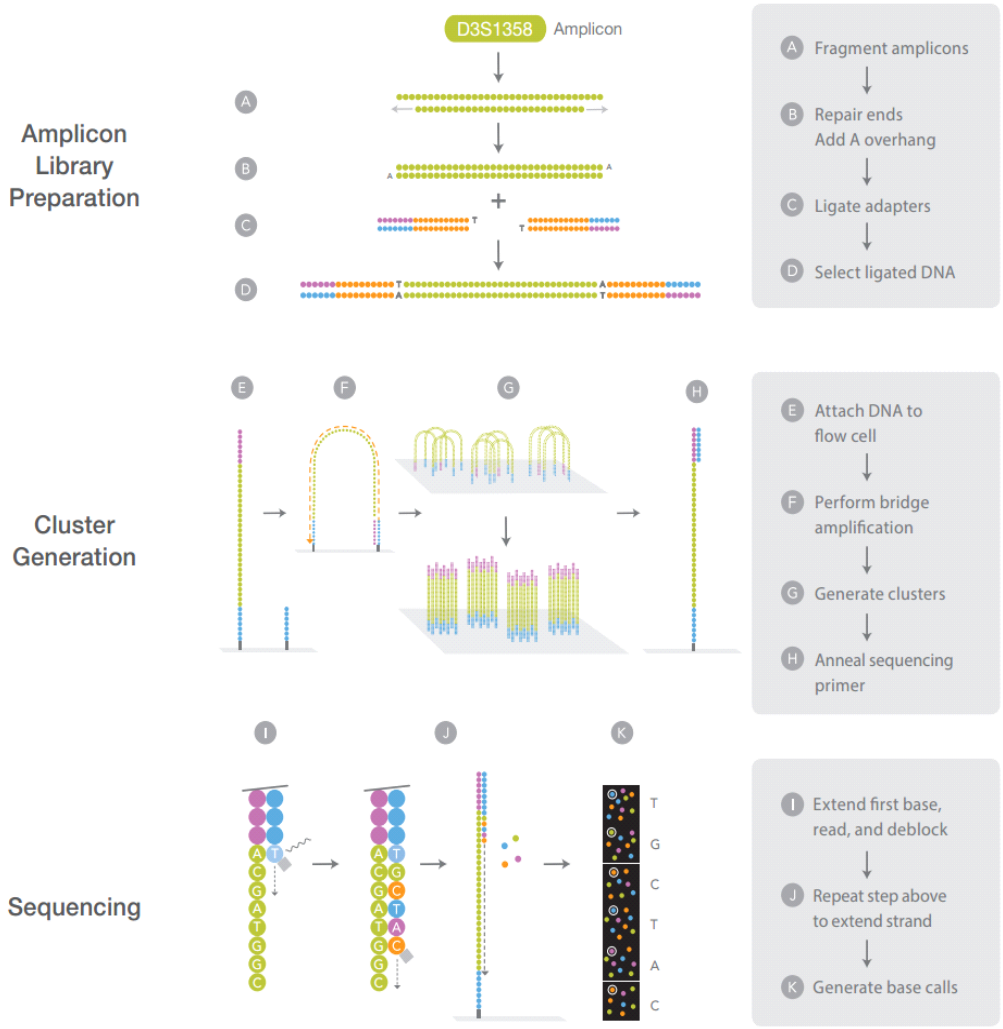


Figure 8-2. Sequencing by Synthesis (SBS) technology workflow (Illumina, 2013b)

8.1.3 Highly multiplexed PCR amplicon sequencing on Illumina MiSeq®

During the last few years, new sequencing platforms have been released such as the Ion Torrent Personal Genome Machine (PGM), the Pacific Biosciences RS (PacBio) and the HiSeq® or MiSeq® system (Illumina), all capable of generating operational sequences (Quail *et al.*, 2012). With respect to Illumina platforms, while the HiSeq 2000 has set the standard for high throughput massively parallel sequencing, Illumina's smallest NGS sequencer, MiSeq® is a lower throughput fast-turnaround instrument aimed for smaller laboratories and the diagnostic market. Therefore, it was considered that the MiSeq® system could be more suitable for the development of forensic-type tests. In particular, while MiSeq® supports various applications, combining Illumina's accurate SBS chemistry with an effective workflow allowing for long 150 bp paired-end reads would be ideal for highly multiplexed PCR amplicon sequencing and the analysis of particular genetic variations.

Illumina has recently focused on potential forensic applications and the development of suitable protocols (Illumina, 2013b). For instance, the Illumina Nextera® XT DNA kit allows for the amplification of forensic loci from 1 ng of starting DNA. Beginning with amplified DNA, the protocol employs a single enzymatic 'tagmentation' reaction to fragment and tag amplicons with sequencing adaptors, making the library compatible with the MiSeq® flow cell. However, PCR amplicons should be larger than 300 bp to ensure even coverage across the length of the DNA fragment as a sequencing coverage about 50 bp from each distal end of a fragment might drop off. To accommodate analysis of more challenging, degraded samples, additional library preparation methods are under development, such as the TruSeq Forensic Amplicon protocol (Illumina).

8.1.4 Illumina MiSeq® for accurate methylation quantification

In this study the ability of NGS based on the MiSeq® system to accurately quantify methylation levels of specific CpG sites will be assessed. Illumina suggests two forms of bisulphite sequencing-based applications that can be employed on a MiSeq® system including whole-genome bisulphite sequencing (WGBS) and reduced representation bisulphite sequencing (RRBS). However, there has been one research group so far that has tested the abilities of MiSeq® in targeted methylation analysis of specific genomic regions (Masser *et al.*, 2013). The authors developed a new approach, termed Bisulphite Amplicon Sequencing (BSAS), for hypothesis-driven and focused, absolute DNA methylation analysis. This method was suggested to be applicable not only to targeted DNA methylation studies but also to confirmation of genome-wide studies, overcoming technical issues faced in predominant approaches like Pyrosequencing®.

Briefly, the BSAS method involves bisulphite conversion of genomic DNA followed by PCR amplification of the regions of interest using target- and converted DNA-specific primers. Libraries are prepared using the transposome mediated (Nextera XT) NGS library preparation technology and the amplified regions are sequenced using the benchtop MiSeq® sequencer. By analysing both rat and mouse DNA methylation standards in replicates, the authors reported a greater precision of methylation quantitation and a 16-fold increase in accuracy compared to Sanger/epigenetic sequencing methylation analysis (ESME) data (Masser *et al.*, 2013) [Figure 8-3]. Also, the variation in methylation quantification was much smaller in BSAS, where there was at most 5% standard deviation.

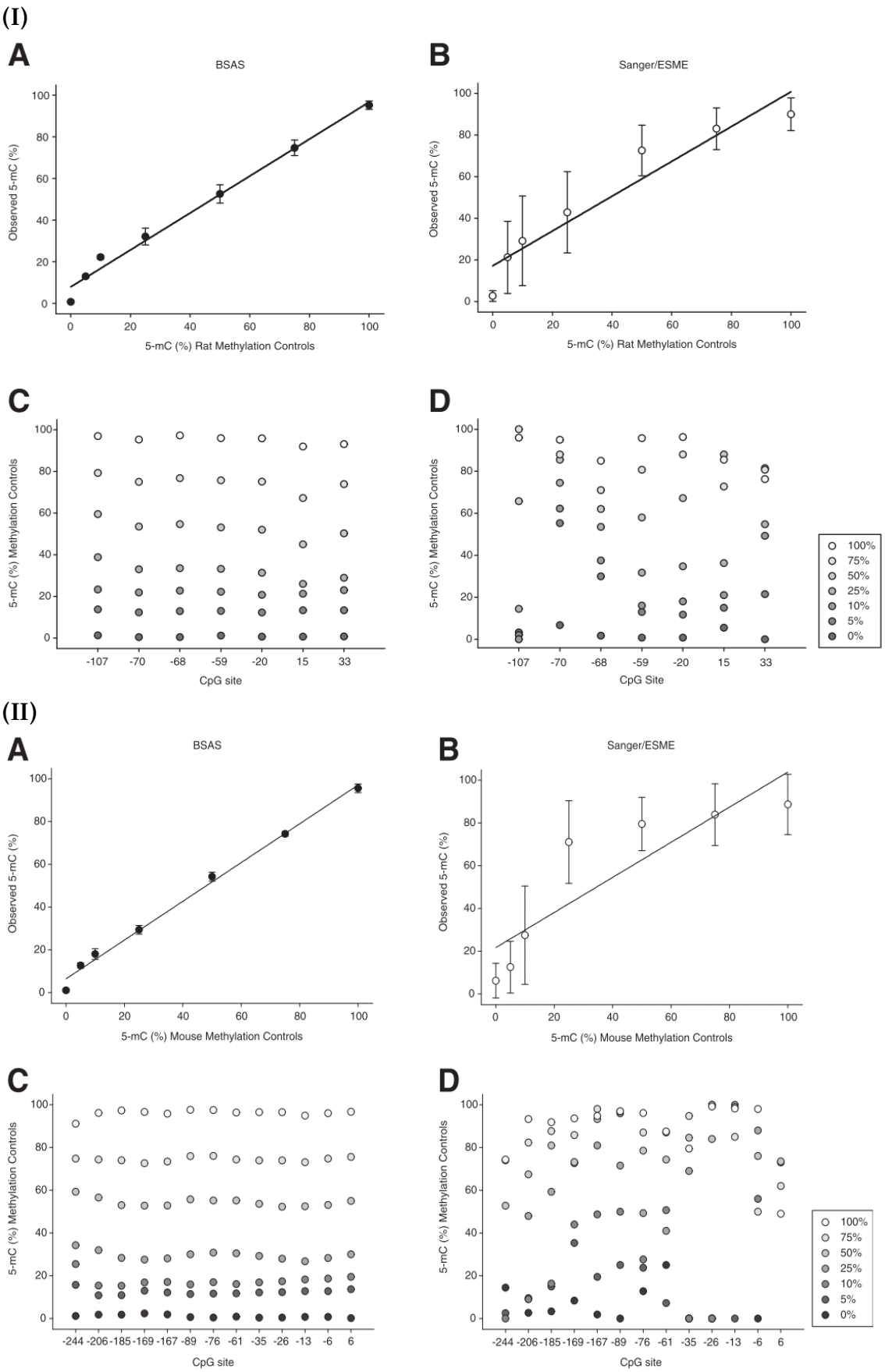


Figure 8-3. BSAS and Sanger/ESME methylation quantification of (I) mouse and (II) rat whole-genome methylation standards (Masser *et al.*, 2013)

(A, B) Observed vs. expected methylation (%) as generated from BSAS and Sanger/ESME methods

(C, D) Average observed methylation for each standard across the CpG sites within analysed amplicon

Based on confidence interval analysis, methylation quantification was not significantly improved at sequencing read depths greater than 1000; therefore, it could be concluded that 1000 reads are sufficient for accurate quantification. While MiSeq® has previously been proposed for high-output CpG validation or evaluation of traditional NGS library construction methods, the authors proposed MiSeq® as a tool for precise orthogonal absolute 5-mC quantitation and validation in this study (Masser *et al.*, 2013). It was also suggested that the MiSeq® system's sensitive optics and more precise base calling permit the sequencing of low-diversity samples. Nevertheless, one has to have in mind potential PCR bias in the original bisulphite PCR as well as bias introduced during library preparation and adapter ligation.

To conclude, next generation sequencing technology has been reported to overcome technical issues of traditional sequencing with regards to throughput, scalability, speed, and resolution. Various epigenetic approaches have been proposed, mostly analysing genome-wide methylation patterns; however, targeted sequencing that would allow for accurately detecting C/T ratios could be achieved. Masser *et al* (2013) reported a bisulphite amplicon sequencing method based on the Illumina's MiSeq® system that enables accurate, high-resolution methylation quantification. Furthermore, since the forensic community already appreciated the advantages of a NGS approach by applying it in various forensic applications such as single nucleotide polymorphisms (SNP) analysis, it was believed that NGS could serve as a better option for validating the proposed age-associated markers included in the age prediction model (Chapter 7).

8.1.5 Aim and Objectives

The purpose of this chapter was to develop an NGS protocol based on the Illumina MiSeq® platform that would potentially allow for age prediction in a forensic scenario through accurate quantification of the methylation levels of the previously proposed age-associated CpG sites.

In order to achieve the aim, the following objectives were met:

- Suitable bisulphite PCR assays for all 16 CpG sites included in the age prediction model were designed and optimised.
- An NGS protocol specifically designed for forensic amplicon analysis was obtained and adjusted for this study.
- The protocol was validated by pre- and post-PCR linearity analysis using DNA standards of known methylation levels.
- The age prediction accuracy was assessed by analysing a total of 34 whole blood samples.
- The reproducibility of the overall method was assessed by studying multiple replicates of the same blood sample.

8.2 Experimental

8.2.1 Blood samples

Whole peripheral blood was collected from 34 healthy volunteers (21 males and 13 females from various ethnic backgrounds) aged 2 to 76 years old (mean age 38.4±21.2 years) [Figure 8-4]. Most blood samples were collected as liquid (200-1,000 μl) and only for the three youngest participants (2-4 years old) a small amount of blood was deposited on a cotton swab. Liquid blood samples were stored at 2-8 °C, while swabs were stored at -20 °C (to minimise DNA degradation) for up to 2 months.

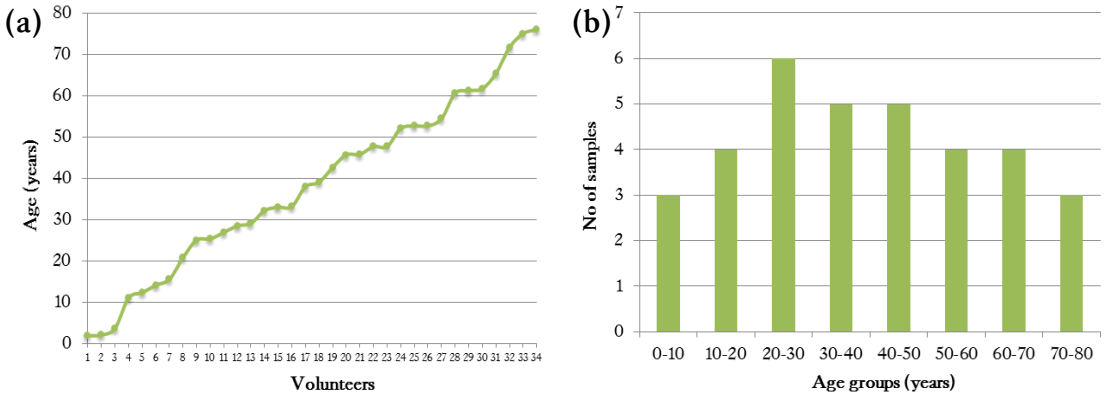


Figure 8-4. Age distribution within blood samples (n=34)
Age distribution of all individuals participated in the study (a) individually and (b) in age groups (years)

8.2.2 Bisulphite PCR assay design

Following the guidelines regarding primer design mentioned in section 2.2.4.1, bisulphite PCRs were designed using the BiSearch primer design tool in order to amplify the genomic regions surrounding the 16 proposed age-associated CpG sites [Table 7-11]. Each assay includes a 10X PCR primer set (forward and reverse) specifically designed to amplify only bisulphite-converted DNA sequences by including converted cytosines in their sequence; however, none of them contains CpG sites to avoid potential bias [Tables 8-1 & 8-2]. Moreover, information on the obtained amplicons including their chromosomal location and number of included bisulphite-conversion controls and CpG sites are presented in Table 8-3, while the DNA sequences of the designed amplicons are shown in Appendix IV.

Table 8-1. Designed bisulphite PCR assays (1-8)

Essential information regarding the designed bisulphite PCR assays are shown, such as the strand that the CpG is located (F for forward and R for reverse), the score assigned by the primer design software (the lower the score the more efficient amplification), the primer sequences and length (forward/reverse), the % G and C content, the melting temperature (T_m), the number of converted cytosines included in the primer sequence (highlighted in red) as well as the length of the final PCR product.

CpG site	DNA strand	Score	Primer Sequence (5'-3')	Length (bp)	% GC	T_m (C°)	Converted Cs	PCR Length (bp)
cg19761273	F	16.71	F TGTTTAGTTTGAAGATTGAG	20	30	54.8	4	150
			R CCTTATTTCCTTTACAAAAA	20	25	56.2	4	
cg27544190	F	20.77	F GGGTAGGATTAAAGTTGA	18	38.9	55.4	0	106
			R CTTAAAAAATAACAATCCCC	19	31.6	55.6	7	
cg03286783	R	41.17	F GTTTTAGTTAGTGGGTG	17	41.2	54	5	181
			R CCCCTCCTCAAATCAAA	17	47.1	56.5	2	
cg01511567	F	18.58	F TATTAGATTTAGTATAGGGG	20	30	54	3	132
			R CCCACAACTATTCAAATA	18	33.3	53.1	1	
cg07158339	R	20.95	F GGAATATGTTTTGTTTAAAA	20	20	54.4	6	122
			R TAATTAACCTCTCTATACCT	20	30	54.6	2	
cg05442902	R	13.61	F GTATGTTTTGGTTTTTGT	18	27.8	53.9	5	109
			R AATAACCTCTAAACTAACC	19	31.6	55	4	
cg24450312	F	24.42	F GTTATTTATAGAGTTTGAG	19	26.3	51.2	5	201
			R TCTACTACAAACCAAA	16	31.2	50.2	5	
cg17274064	R	20.72	F AGGGAATAAGTATTTTTTT	18	22.2	53.2	4	139
			R CTCACAATCAAACTTCTATATAC	23	30.4	56	4	

Table 8-2. Designed bisulphite PCR assays (9-16)

Essential information regarding the designed bisulphite PCR assays are shown, such as the strand that the CpG is located (F for forward and R for reverse), the score assigned by the primer design software, the primer sequences and length (forward/reverse), the % G and C content, the melting temperature (T_m), the number of converted cytosines included in the primer sequence (highlighted in red) as well as the length of the final PCR product.

CpG site	DNA strand	Score		Primer Sequence (5'-3')	Length (bp)	% GC	T_m (C°)	Converted Cs	PCR Length (bp)
cg02085507	R	15.01	F	GTTAATGGATTTGGTTTTG	19	31.6	55.3	4	186
			R	AACTCAAAAAAATCCTTCCT	19	31.6	56.8	2	
cg20692569	F	31.34	F	TTGTTGTTGTGGTAGT	16	37.5	51.8	5	160
			R	AACCCAACAAATTAAA	16	25	51.8	8	
cg04528819	R	13.14	F	AATAGGTTTTGGTGTAGTT	19	31.6	56	4	138
			R	CAACCTCTAATAAATTCTCT	20	30	54.6	6	
cg08370996	R	29.24	F	GTGTTAAAGTTTATTATATAGA	22	18.2	53.3	3	187
			R	AAAAAAAAAAACACACAC	18	22.2	54.1	3	
cg04084157	F	20.49	F	GAGGGTGTTTGTTTTTTTT	18	33.3	56.6	7	111
			R	AACATTTCATTCATTCATTC	20	25	53.1	1	
cg22736354	F	16.33	F	GTTGAGTTTTAGGAGTTTTTAT	20	30	55.8	6	201
			R	CTTTAAAAAAATTTAACCACC	20	25	56.4	5	
cg06493994	R	20.26	F	GGAGAGTAAGTTAAGAAATA	20	30	54.7	2	150
			R	AACCTACCAAAAACCAAC	18	38.9	57.5	5	
cg02479575	F	16.08	F	GGAGGAGAATGTTATTTATT	20	30	55.1	1	143
			R	CTATCCAAAATTCTAAAAAC	20	25	54.4	7	

Table 8-3. MiSeq® DNA methylation assays (16)

	Assay	Conversion controls	CpG sites	Chromosomal location to be sequenced
1	cg19761273	42	2	17: 80,232,017-80,232,166
2	cg27544190	18	4	21: 33,785,414-33.785.519
3	cg03286783	40	10	15: 44,580,864-44,581,044
4	cg01511567	12	3	11: 57,103,582-57,103,713
5	cg07158339	23	1	9: 71,650,150-71,650,271
6	cg05442902	32	2	22: 21,368,989-21,369,097
7	cg24450312	66	27	1: 206,681,003-206,681,203
8	cg17274064	23	2	21: 40,033,806-40,033,944
9	cg02085507	38	12	19: 6,739,164-6,739,349
10	cg20692569	37	20	7: 72,848,365-72,848,524
11	cg04528819	35	8	7: 130,418,073-130,418,210
12	cg08370996	41	13	15: 96,873,886-96,874,072
13	cg04084157	19	8	7: 100,808,988-100,809,098
14	cg22736354	47	18	6: 18,122,551-18,122,751
15	cg06493994	33	11	6: 25,652,542-25,652,691
16	cg02479575	31	10	19: 4,769,688-4,769,830

8.2.3 Sample analysis

Genomic DNA from 200 µl of each blood sample or the whole swab was extracted using the EZ1 Blood DNA kit as described in section 2.2.1.3 and eluted in either 200 µl (liquid) or 50 µl (swabs) of elution buffer. DNA samples were then quantified in duplicate using the Quantifiler Human DNA Quantification kit as described in section 2.2.2.1. Afterwards, 100 ng of extracted DNA or each of the DNA methylation standard was treated with sodium bisulphite using the MethylEdge™ Bisulphite Conversion System (Promega) as described in section 2.2.3.3; bisulphite-converted DNA was eluted in 40 µl of elution buffer and stored at 2-8 ºC for up to one week.

All samples were amplified using the ZymoTaq™ premix (Zymo Research) and following the optimised conditions of each PCR assay. In order to reduce cost, half-volume PCR reactions were used throughout analysis. Briefly, each PCR reaction consisted of 6.25 μl of ZymoTaqPreMix, 0.75 μl of 25 mM $MgCl_2$ for a final concentration of 3.2 mM (since the ZymoTaq™ Premix also contains 1.75 mM $MgCl_2$), 0.5 μl of each PCR primer (for a final concentration of 0.4 μM), 1 μl of bisulphite DNA template and 4 μl of nuclease-free water, for a total reaction volume of 13 μl. The thermocycling program used was: 95 °C for 10 minutes, followed by 30 or 45 cycles of 94 °C for 30 seconds, T_m for 30 seconds, 72 °C for 30 seconds, and a final extension step of 72 °C for 7 minutes. The optimised T_m was as follows: 48 °C for cg07158339, cg17274064, cg02085507, cg20692569 and cg02479575, 50 °C for cg19761273, cg27544190, cg01511567, cg24450312 and cg04528819 and 52 °C for cg03286783, cg05442902, cg08370996, cg04084157, cg22736354, cg06493994.

Finally, for each sample 2 μl of each PCR product were pooled for a total volume of 32 μl. In particular, to ensure that sufficient amounts were present for downstream analysis, 10 μl of each PCR product for the 0% and 100% methylation DNA controls were pooled instead (total volume=160 μl). Post-PCR methylation controls were made up by mixing appropriate amounts of the unmethylated (0%) and methylated (100%) pooled PCR products as shown in Table 8-4. Subsequently, final pooled PCR products were sequenced using the TruSeq Forensic Amplicon protocol (Illumina). Information regarding all different methodological steps included in this approach as well as comprehensive experimental conditions are extensively described in section 2.2.6.2; any changes will be stated otherwise.

Table 8-4. Composition of post-PCR linearity DNA methylation controls

Post-PCR linearity controls	Pooled PCR product (μl)	
	0% control	100% control
0%	31	0
10%	27.9	3.1
20%	24.8	6.2
30%	21.7	9.3
50%	15.5	15.5
70%	9.3	21.7
80%	6.2	24.8
90%	3.1	27.9
100%	0	31

8.3 Results

8.3.1 Optimisation of bisulphite PCR assays

As mentioned in Chapters 5 and 7, bisulphite-treated DNA is considered as one of the most challenging templates to amplify due to its less complex nature (T-rich). Therefore it was necessary that PCR reactions for each marker were well optimised before downstream analysis. Each assay was optimised using an annealing temperature gradient, various concentrations of $MgCl_2$ and primer as well as different PCR cycling conditions. Figure 8-5 shows the final step of the optimisation process where a set of three annealing temperatures (48 °C, 50 °C and 52 °C) were tested for each assay. As shown, there was a variation in PCR efficiency, with cg17274064 and cg20692569 being the weakest markers. Since cg20692569 has one of the 'worst' scores given by the primer design software, it is believed the low PCR yield could be due to its 'difficult', CpG-rich sequence.

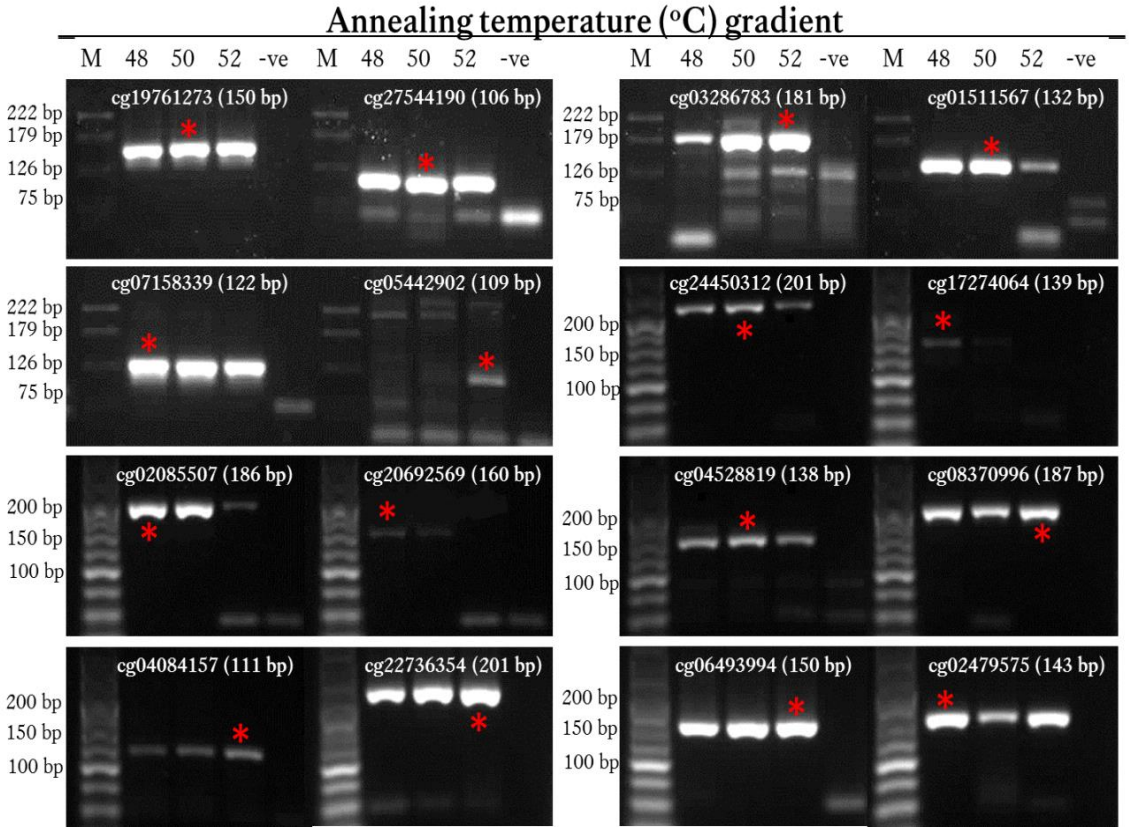


Figure 8-5. Final optimisation for all 16 designed bisulphite PCR assays (45 PCR cycles)
Agarose gel images show the results of the final optimisation step for all PCR amplicons (expected length in brackets) regarding annealing temperature. For each assay, the first column represents the DNA marker (M), the next three columns show the resulting amplification bands for temperatures 48 °C, 50 °C and 52 °C respectively, while the last column represent the PCR negative (no-template) control which was amplified at 50 °C. Red stars indicate the final selected temperature for each assay.

For most assays no non-specific amplicons were present; however the optimisation of cg27544190, cg03286783 and cg05442902 was challenging due to unwanted products resulting from primer interactions. Nevertheless, the signal of the desired amplicons was much stronger than the non-specific one, so the proposed PCR conditions were accepted; the MiSeq® will only use the reads from the expected DNA sequence for analysis. Also, although primers had already been designed for cg05442902 as part of the bisulphite Pyrosequencing® analysis in Chapter 7, primers were redesigned in an attempt to increase its low PCR efficiency [Figure 7-3]. For certain assays such as cg07158339 all temperatures seemed to work well, therefore one temperature was chosen arbitrarily; red stars in Figure 8-5 indicate the selected annealing temperature for each assay. Since all assays share similar PCR conditions, it might be feasible to amplify them in a multiplex reaction. For this project, singleplex PCRs were used in order to avoid potential marker-to-marker PCR amplicon imbalance, but developing a multiplex PCR reaction could be assessed in the future.

8.3.2 Validation of methylation quantification

To test the performance of the method as well as assess the accuracy of quantitative methylation analysis, seven enzymatically-generated whole-genome human methylation controls at mixed ratios (0%, 5%, 10%, 25%, 50% and 100%) were analysed in duplicate for CpG methylation levels using the proposed protocol.

8.3.2.1 Assessment of pooled PCR amplicons

A total of 100 ng of each control were bisulphite-treated and eluted in 40 µl of elution buffer. Since Promega suggests that there is likely a ~20% loss of DNA during treatment, the final bisulphite-converted DNA solutions were expected to have a concentration of ~2 ng/µl; therefore, a total of ~2 ng of DNA template was added in each PCR reaction. Furthermore, to assess amplification bias in designed bisulphite PCRs, methylation controls were amplified for 30 and 45 cycles to assess the effect of PCR cycles upon linearity of methylation quantification. Using the PCR optimisation results as a guide for individual PCR efficiency [Figure 8-5], 1 to 5 µl of each individual PCR product (16 in total) were pooled for a final volume of 30 µl. Pooled PCR products were assessed using the Qubit dsDNA HS assay as described in section 2.2.6.2.2 and the results are shown in Figure 8-6.

Methylation controls	Qubit reading (ng/ml)					
	30 cycles			45 cycles		
	1	2	Average	1	2	Average
0	6.01	5.8	5.905	55.3	51.7	53.5
0.05	6.01	5.9	5.955	58.8	58	58.4
0.1	5.71	5.76	5.735	58.2	57.4	57.8
0.25	6.32	6.24	6.28	72.7	72.3	72.5
0.5	6.1	6.01	6.055	63.8	63.1	63.45
0.75	6.25	6.15	6.2	70.9	69.8	70.35
1	6.05	5.91	5.98	68.6	68.1	68.35
Mean			**6.02**			**63.48**
Standard deviation			**0.18**			**7.19**

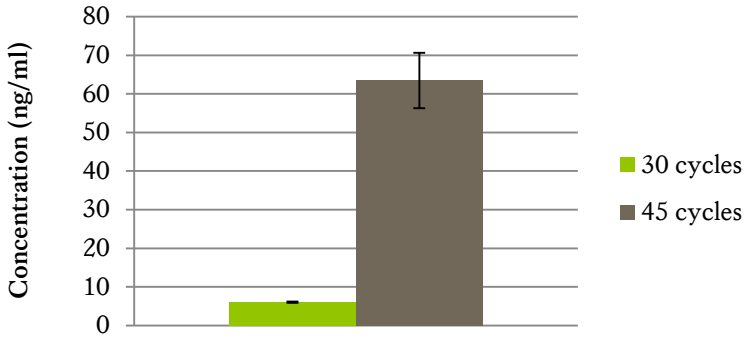


Figure 8-6. Concentration readings of pooled PCR products after 30 and 45 PCR cycles
Pooled PCR products for all seven methylation controls produced by either 30 or 45 PCR cycles were quantified using the Qubit dsDNA HS assay (Life Technologies). All samples were analysed in duplicate.

As shown in Figure 8-6, all controls gave similar readings to one another when using 30 or 45 PCR cycles. However, the extra 15 cycles resulted in a 10-fold increase in concentration as expected (30 cycles –1,203 ng/ml, 45 cycles – 12,696 ng/ml). Since Illumina suggests a 20-2000 pg/µl DNA input, for the first set of controls (30 cycles) all 30 µl was used for library preparation (average=722 pg/µl), while for the second set (45 cycles) only 4 µl of pooled PCR products were used (856-1160 pg/µl).

8.3.2.2 Evaluation of generated libraries

Once the library preparation was complete, libraries were evaluated using a qPCR library quantification kit (KAPA). The average obtained concentration was 397±239 nM. There was significant variation in the final yield of libraries, which is suspected to be due to the extensive tube manipulation during preparation. The numerous washes using the magnetic beads could also introduce variation; however these quantities were sufficient since final library requirements were only 6-20 pM. Only two controls, namely the 25% and 50% methylation standards amplified using 45 cycles yielded very low-quality libraries (7.83 and 8.8 nM respectively). Since the corresponding

PCR product concentrations were high, this could be due to experimental error. Excluding these samples, comparing the 30- and 45-PCR cycle libraries there was no statistically significant difference between them (p=0.98).

8.3.2.3 Quality control of MiSeq® run

In this first attempt, the MiSeq® run was set up to perform 110 cycles for each sequencing direction, which indicates how many nucleotides are added in total from each direction of sequencing. Since all PCR amplicons are between 106-201 bp long, 110 cycles were considered sufficient; however as a result a decrease of read numbers at the ends of each sequence was observed. As shown in Figure 8-7a, 91.5% of total reads passed the MiSeq® quality control (Q>30); 96.1% in the first read and 87.4% in the second one. Also, the flow cell demonstrated a cluster density of 985 K/mm^{2}that closely matches the desired 1000 K/mm^2 recommended by Illumina [Figure 8-7b]. There were a total of 19 million reads, 89.86% of which were successfully identified. The contribution of each sample to the total identified reads was an average of 3.2% (standard deviation=0.96%). From this pool, almost all (98.9%) had less than three errors (number of usable cycles); therefore the total error rate was as low as 0.26%. Furthermore, in all cases bisulphite conversion rates were confirmed to be >98% by analyzing the cytosine to thymine ratio for non-CpG cytosine positions.

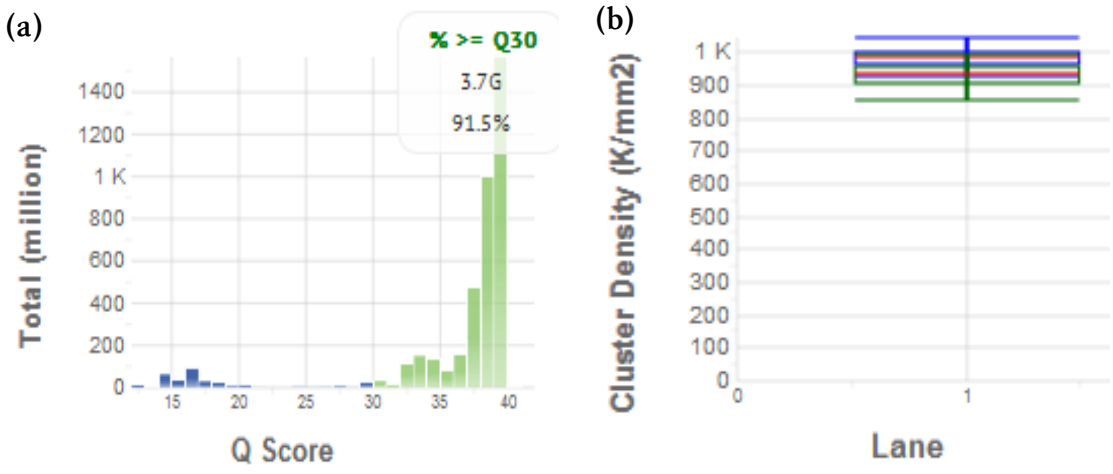


Figure 8-7. MiSeq® run quality control parameters
(a) Distribution of Q score among samples, with 91.5% passing the instrument's quality control (Q30)
(b) Average cluster density (K/mm^2) of the flow cell. Blue bars indicate the 'ideal' 1000 K/mm^2, while the green are the observed one (985 K/mm^2)

8.3.2.4 Distribution among samples and amplicons

As illustrated in Table 8-3, each sequenced DNA fragment contained between 1 to 27 CpG sites; however for this analysis only data from the CpG sites in question (16 in total) were used. The total reads of each sample were calculated by adding each marker's individual read number (16 CpGs) while the average reads for each CpG site was also obtained using data for all methylation controls. Read numbers ranged between 492,201 to 1,159,547 and there was a good distribution between them [Figure 8-8]; however the 25% and 50% methylation standards for the 45 cycles set failed to produce sufficient reads. Since these two samples had resulted in very low-quality libraries before, it is assumed that the failure was not due to sequencing issues.

On the other hand, the distribution of individual assays was not as even as desired [Figure 8-9]. While some markers resulted in large numbers of reads (e.g. cg05442902 constantly gave >100,000 reads), there were some that had relatively few reads (e.g. average of 2,303 reads for cg17274064). This could be due to PCR efficiency issues; for example, cg17274064 and cg20692569 had demonstrated weaker bands in Figure 8-5. Also, errors during pooling the individual PCR amplicons could have been introduced since 1 to 5 μl of each PCR product were pooled; however, this deviation could potentially be avoided if the individual PCR amplicons are quantified before pooling. Lastly, variations caused during library preparation could also be the reason of this deviation, since the generated libraries had been normalized, yet diluted hundreds of times.

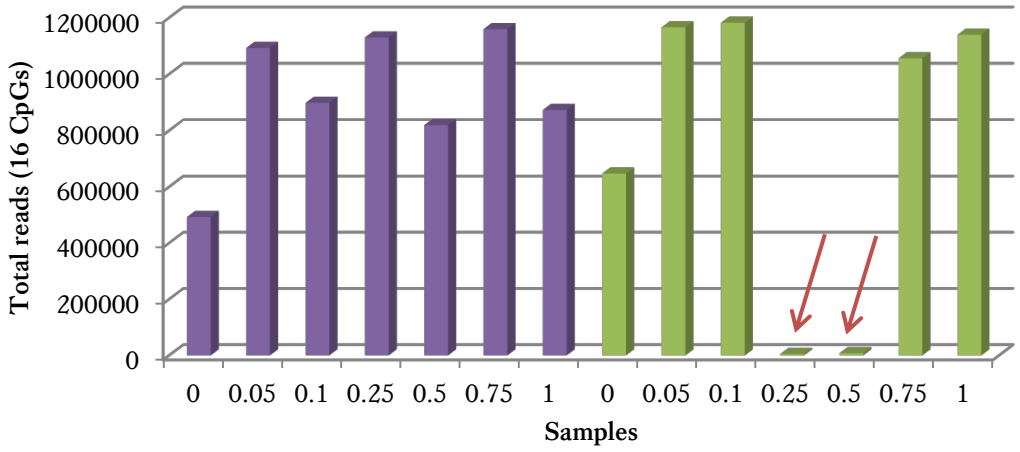


Figure 8-8. Total reads for each methylation standard as obtained for all 16 CpG sites
Methylation controls had been amplified either using 30 (purple) or 45 (green) cycles. Red arrows indicate the two samples that 'failed'.

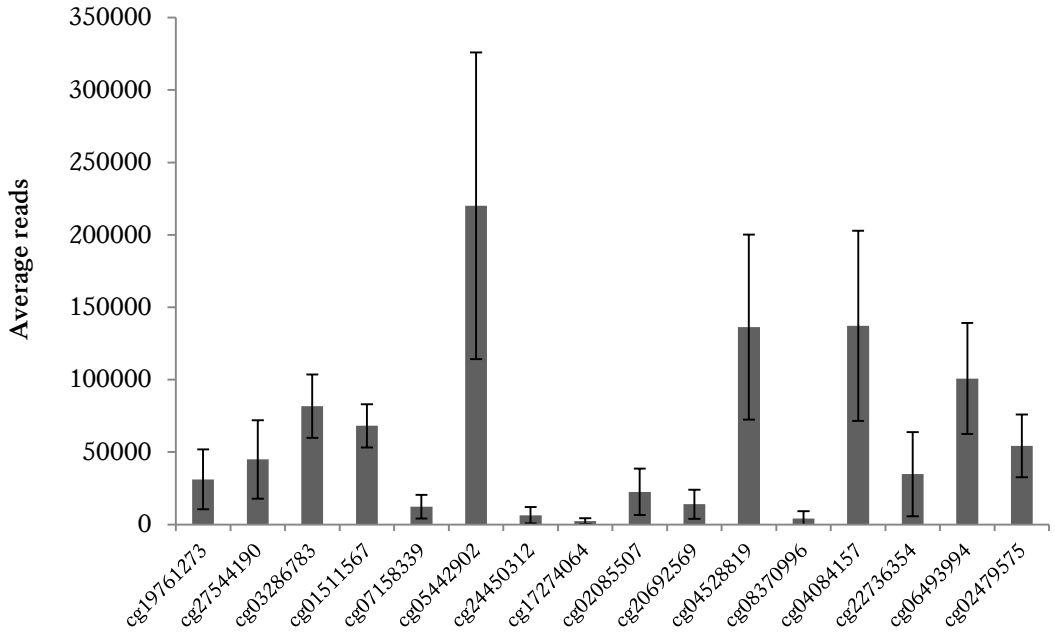


Figure 8-9. Average read number per assay (16)
The average reads used in this graph were calculated by using all methylation standards except the two samples that resulted in very low reads. As shown, there was a large difference among the different DNA fragments which could be due to differences in PCR efficiency or variations during pooling and library preparation. Error bars correspond to standard deviation values.

8.3.2.5 Pre-PCR linearity of methylation quantification

To assess how linear the methylation quantification was, standard curves were generated from each set of methylation controls (amplified by 30 and 45 PCR cycles) either by taking the mean methylation level across all measured CpG sites (151 CpGs in total) [Figure 8-10a, b] or by averaging the obtained methylation levels of only the 16 CpG sites of interest [Figure 8-10c, d]. It should be noted that the 25% and 50% controls for the 45-cycle set were not included as they generated very low reads. For the first set of standards (30 cycles), both standard curves were able to fit linear lines (R^2=0.992), while for the second set (45 cycles) the observed linear correlation was slightly weaker (R^2=0.973 and R^2=0.982 respectively). Generally, for all sets of controls the methylation quantification was accurate as evidenced from the high correlation coefficients (r=0.99 in all four graphs); however the precision was greater for the standards amplified only using 30 cycles, more likely due to the weaker effect of amplification bias during PCR. The latter is more evident when looking at each CpG site individually (data not shown). It is believed that the extra 15 PCR cycles introduce further amplification bias especially in the low methylated samples. It was decided that the first set (30 cycles) would be used for further analysis.

Moreover, it was noted that using only the methylation data from the 16 CpG sites did not significantly change the quantification accuracy compared to when using all 151 CpGs [Figure 8-10a, c]. In fact, for the second set of controls (45 cycles), the linear correlation between observed and detected methylation was slightly improved. On the other hand, as indicated by the error bars (standard deviation) the variation in detected methylation was still considerably high and did not match the reported maximum 5% error in the study by Masser et al. (2013). Nevertheless, the small overestimation of methylation observed in the 5%-25% standards had also been detected in the Masser study.

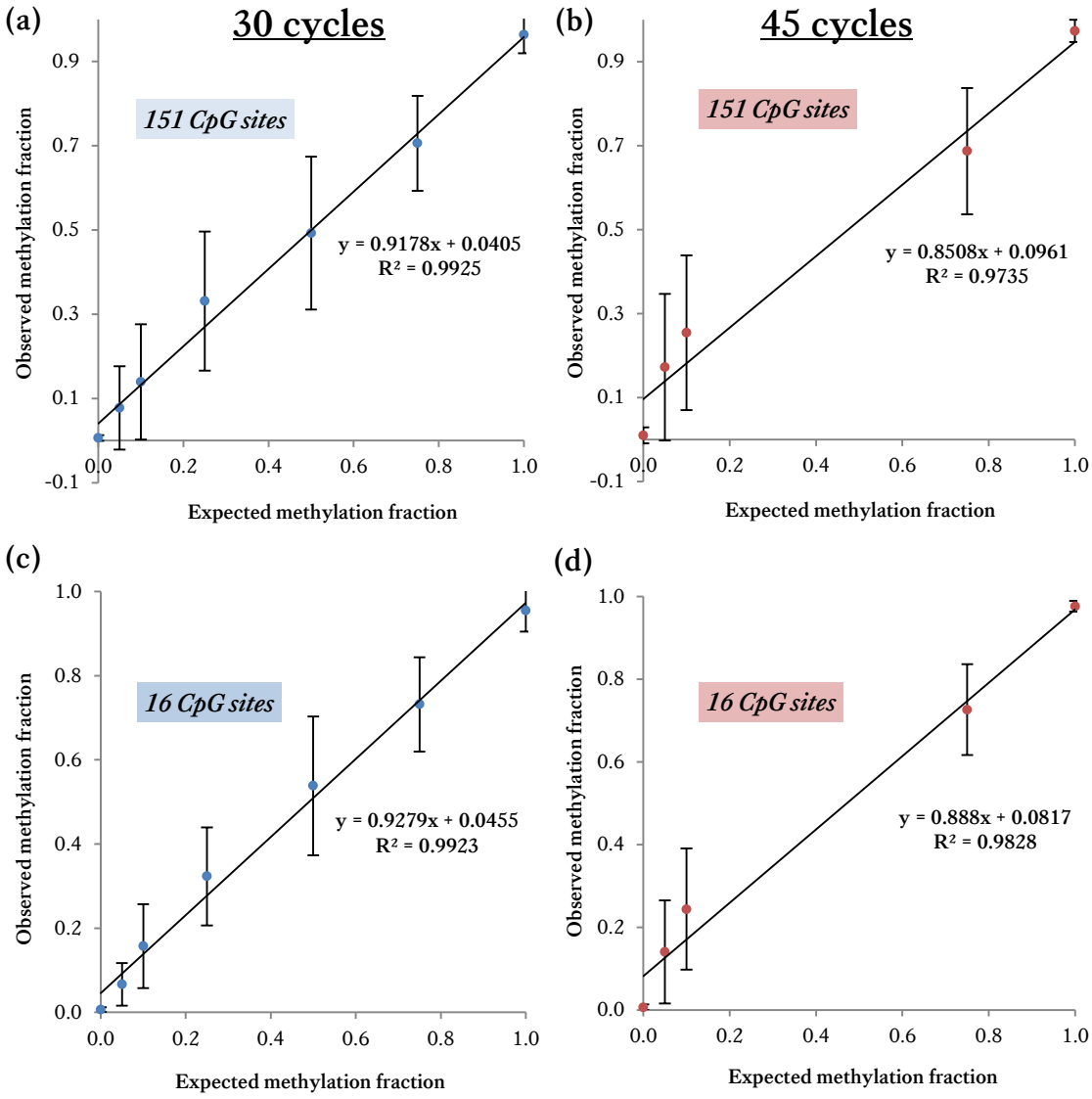


Figure 8-10. Standard curves generated by 30- (a, c) and 45-cycle (b, d) PCR amplicons
The observed quantified methylation was plotted against the expected methylation fraction for all standards. Points represent the mean of either 151 (a,b) or 16 (c,d) CpG sites analysed for each methylation standard. Error bars correspond to standard deviation.

By using genome-wide methylation standards to generate methylation fraction standard curves, it was demonstrated that the proposed method was capable of accurately quantify CpG methylation but significant variation was observed among individual markers. The graphs above were made under the assumption that the standards used had the 'correct' methylation levels. According to the manufacturer (EpigenDx), the low methylated standard (0%) is actually less than 5%, while the highly methylated one is definitely above 85%. Considering that these two standards were used to prepare the remaining standards (5%, 10%, 25%, 50% and 75%), it was thought that the methylation values could be slightly different from the expected. In an attempt to eliminate errors, it was decided that the expected values should be 'normalised'. Using the average methylation detection for the low and highly methylated controls obtained by both sets of controls, the expected methylation values were calculated for the rest of the controls. The individual 'normalised' observed *vs*. expected methylation graphs of each CpG site (30 cycles) are shown in Appendix V. Most standard curves fit linear lines; however quadratic or cubic polynomial curves provided better fits for some markers resulting in improved correlation. In most cases, it seemed that the unmethylated or methylated allele was favored during amplification therefore resulting in underestimation or overestimation of methylation respectively. Figure 8-11 shows four different examples; (a) marker cg19761273 showed preferential quantification of the methylated allele, (b) in marker cg06493994 the unmethylated allele was preferred, (c) marker cg04084157 resulted in linear quantification and (d) marker cg20692569 showed a more complicated pattern.

Similar results regarding non-linear amplification of bisulphite-treated DNA have previously been reported in the literature (Fernandez-Jimenez *et al.*, 2012; Moskalev *et al.*, 2011). To date, no common experimental approach has been suggested to overcome this problem; therefore scientists have applied effective methods of 'correcting' these types of bias instead. Once methylation standards are analysed and the observed deviation from the expected methylation values is plotted, then the equation of the best-fitting regression curve can be used for correction. This process could be applied to any CpG site of interest without the need for time-consuming PCR optimization. The equations used to correct the observed biased methylation data in this study are shown in Table 8-5.

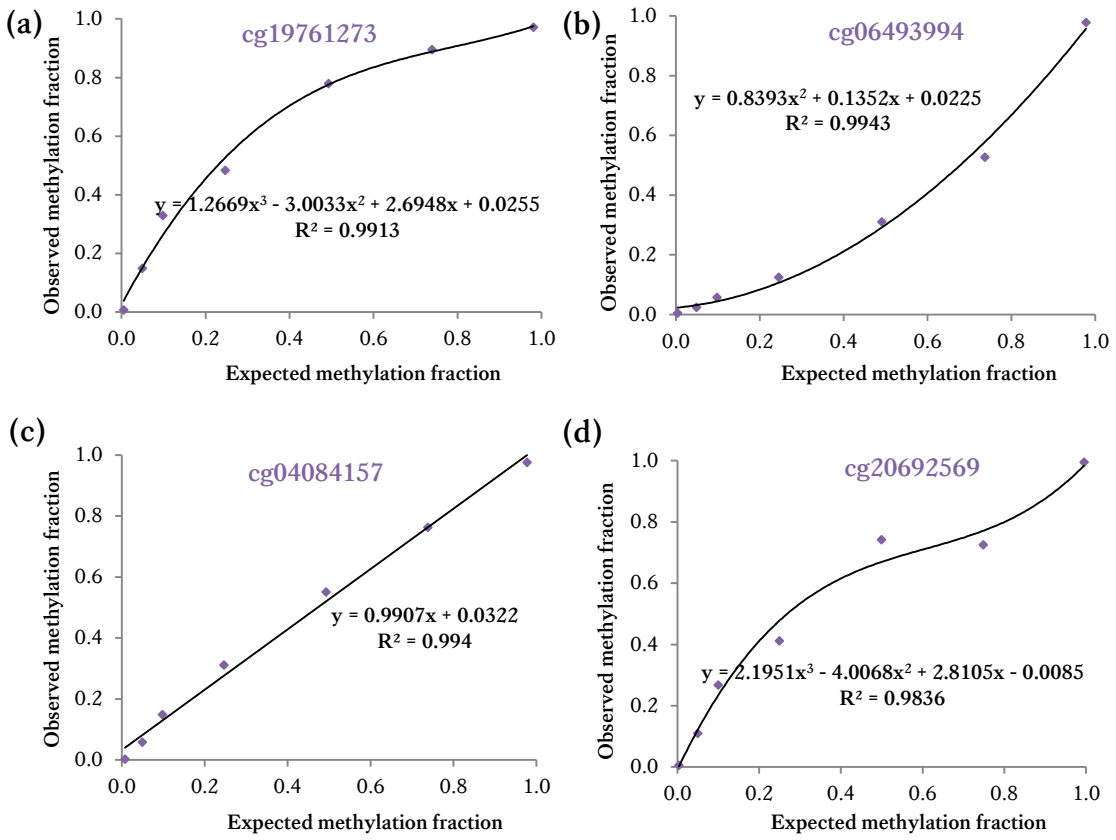


Figure 8-11. Different patterns of methylation quantification observed in four CpG sites
(a) cg19761273 showing preferential quantification of the methylated allele, (b) cg06493994 showing preferential quantification of the unmethylated allele, (c) cg04084157 demonstrating linear quantification and (d) cg20692569 showing a more complicated pattern where low methylated samples are overestimated while highly methylated samples are underestimated.

Table 8-5. Equations obtained from the best-fitting regression line for all 16 CpG sites

16 CpG sites	R^2 value	Equation (y=observed, x=expected)
cg19761273	0.9913	$y = 1.2669x^3 - 3.0033x^2 + 2.6948x + 0.0255$
cg27544190	0.9968	$y = -0.3803x^2 + 1.3179x + 0.0215$
cg03286783	0.9934	$y = 0.5103x^2 + 0.4824x + 0.006$
cg01511567	0.9983	$y = -0.5805x^2 + 1.5621x - 0.0014$
cg07158339	0.9939	$y = 1.1725x^3 - 2.605x^2 + 2.4224x - 0.0215$
cg05442902	0.9906	$y = 1.5036x^3 - 3.2402x^2 + 2.7272x + 0.0022$
cg24450312	0.9758	$y = 1.2559x^3 - 1.9871x^2 + 1.7805x - 0.051$
cg17274064	0.9912	$y = 0.6402x^3 - 1.4133x^2 + 1.7862x - 0.0209$
cg02085507	0.9905	$y = 0.6715x^3 - 1.2861x^2 + 1.6257x + 0.007$
cg20692569	0.9836	$y = 2.1951x^3 - 4.0068x^2 + 2.8105x - 0.0085$
cg04528819	0.9903	$y = 0.8618x^3 - 0.9387x^2 + 1.0871x + 0.0317$
cg08370996	0.9673	$y = 1.1402x^3 - 2.5073x^2 + 2.3218x + 0.0038$
cg04084157	0.9940	$y = 0.9907x + 0.0322$
cg22736354	0.9855	$y = 0.1132x^2 + 0.7958x - 0.0086$
cg06493994	0.9943	$y = 0.8393x^2 + 0.1352x + 0.0225$
cg02479575	0.9979	$y = 0.8236x^2 + 0.1745x + 0.0138$

8.3.2.6 Post-PCR linearity of methylation quantification

To further validate the method and potentially identify the source of bias, nine methylation controls at mixed ratios (~10% intervals) were prepared post-PCR and analysed in a second run. The 0% and 100% methylated controls were amplified with all assays using the same experimental conditions as before and the concentration of each individual assay was measured. Only small differences amongst the two PCR products for each assay were observed so 2 μl of each were pooled into the final pooled amplicons. It was calculated that the unmethylated control contained a total of 256 ng in 10 μl, while 10 μl of the methylated control were comprised of 331 ng. Since all experimental conditions were the same, it can be assumed the methylated allele was amplified with higher efficiency. Considering this difference, the rest of the controls were generated as described in section 8.3.2; their exact methylation levels were also calculated and normalised (0.6%, 11.8%, 22.9%, 33.5%, 52.9%, 70.5%, 78.7%, 86.5% and 93.9%).

The standard curve of the average observed vs. expected methylation fitted a linear line (R^2=0.991) and did not change significantly compared to the pre-PCR linearity (data not shown). However, when looking at CpG sites individually, it was observed that for some of them the quantification turned to linear (cg27544190, cg03286783, cg04528819) whereas for some others the bias was not eliminated (cg19761273, cg08370996, cg22736354). These results could further support that the 0% and 100% methylated controls are amplified with different rates, bias that could not abolished through the proposed normalization. Yet, they could also indicate that potential bias occur post-PCR during the sequencing process.

8.3.2.7 Cross-evaluation of cg05442902 between sequencing platforms

As previously mentioned, in the final set of age-associated CpG sites selected for validation via next generation sequencing, there was one marker (cg05442902) that had previously been validated using Pyrosequencing® (section 7.3.1.4). Therefore, it was interesting to compare the accuracy of methylation quantification using the same set of standards with pre-defined methylation status. The primer sets used were not the same; however conclusions concerning both instruments' capability to accurately measure methylation levels could be assessed. As shown in Figure 8-12, although both

techniques generated a high correlation coefficient (r=0.96), Pyrosequencing® failed to detect higher than 60% methylation indicating that possible amplification bias could have a greater effect in this method. One should note that a standard bisulphite Pyrosequencing® protocol includes a total of 45 PCR cycles.

	Methylation fraction						
cg05442902	0.01	0.05	0.10	0.25	0.49	0.74	0.97
MiSeq®	0.02	0.10	0.31	0.46	0.75	0.85	0.97
Pyro®	0.05	0.13	0.21	0.28	0.46	0.56	0.59

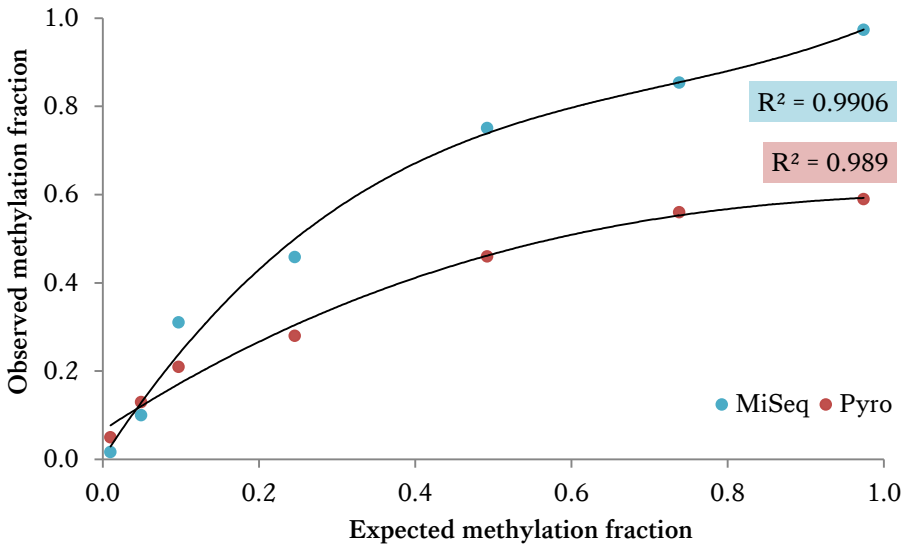


Figure 8-12. Comparison of quantification accuracy between sequencing platforms
The same set of methylation standards were analysed by both Pyrosequencing® and MiSeq® and their observed methylation values were plotted against the expected 'normalised' ones. As shown, MiSeq® demonstrates higher accuracy. The best-fitted regression line for both methylation data point sets was a cubic polynomial curve.

This first validation showed that the proposed method based on MiSeq® was capable of generating high quality libraries and accurately quantifying methylation levels. However, it was believed that the accuracy could be further improved based on key findings. To obtain accurate methylation values, using 30 PCR cycles was preferable; not only did it result in sufficient amounts of amplicons that could be used directly for library preparation, but also resulted in smaller variation and greater accuracy compared to 45 PCR cycles. However, a correction of detected amplification bias is required and should be performed for each marker individually. Also, the strategy when pooling the various PCR products could be adjusted so that better distribution among markers is achieved. Although the markers showing low PCR efficiency during PCR optimisation resulted in fewer reads, there was a significant variation of marker distribution among samples, more likely due to error during the normalization step.

8.3.3 Implementation in blood for age prediction

To test the method's ability to predict age, a set of 34 blood samples were tested. Analysing six replicates of one blood sample also assessed the reproducibility of methylation quantification.

8.3.3.1 Blood DNA yield

As mentioned in the experimental section, for this study a total of 34 blood samples from individuals aged 2-76 years old were collected. The samples were either stored in liquid form or spotted onto cotton swabs for up to 2 months before analysis. Following isolation of DNA using 200 µl or the whole swab of blood, samples were quantified in duplicate using Quantifiler (Life Technologies). Regarding the liquid blood samples, the average measured concentration was 40.71±12.66 ng/µl which corresponds to a total DNA yield ranging between 3.31-12.6 µg (average=7.53 µg). In some cases, the extraction method proved to be even more efficient than the manufacturer suggested (4-8 µg). Since only 100 ng of DNA was needed for bisulphite conversion and further analysis, we could assume that 2-5 µl of blood would be sufficient as starting material. However, no sensitivity analysis was performed in this study but could be assessed in the future. Additionally, as shown in Figure 8-13 it seemed that storage time (6-70 days) at 4 °C did not have any effect on total DNA yield.

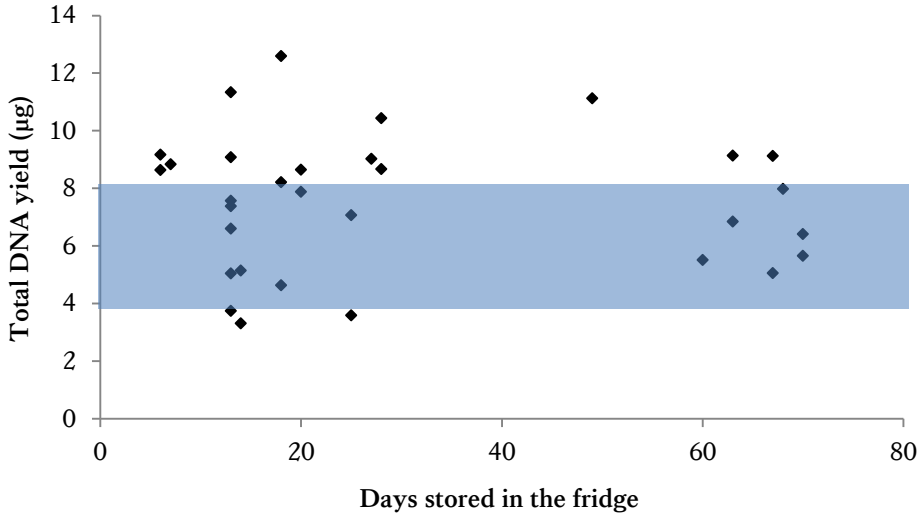


Figure 8-13. Effect of storage time in total DNA yield
The total DNA yield for each blood sample plotted against the number of days that was stored at 4 °C. The blue bar represented the expected DNA yield suggested by the manufacturer when using 200 µl of blood for extraction and following the same experimental conditions.

8.3.3.2 PCR product pooling strategy

In the previous section when analysing DNA standards with pre-defined methylation levels, PCR optimisation results were used as a guide before pooling individual PCR products [Figure 8-5]; therefore 1 to 5 μl of each PCR product were pooled depending on PCR efficiency. However, even though the concentration among pooled amplicons was quite similar [Figure 8-6], there was a significant variation in read numbers among markers [Figure 8-9]. Consequently, it was believed that these differences could be avoided by adjusting the pooling strategy via quantifying the singleplex PCR reactions prior to pooling. The latter was not feasible to be done for all samples since there were more than 700 reactions in total. Nevertheless, even though the employed method (Qubit dsDNA HS assay) also quantifies primer-dimers and secondary structures of unused bisulphite DNA strands, therefore it does not provide very accurate results, individual PCR products were quantified for one blood sample in duplicate.

As illustrated in Figure 8-14 (green column), all individual PCR assays seemed to have resulted in successful amplification with concentrations ranging from 1.3 to 2.6 ng/μl. If 2 μl of each marker were pooled together for this sample, the total amount of PCR amplicons would be 56.31 ng resulting in a final concentration of 1.77 ng/μl. Since the reported variation of different concentrations was not high, it was concluded that this would be a good pooling strategy. Therefore, for all samples, 2 μl of each individual PCR product were pooled to form the final amplicon pool, which was then quantified using Qubit [Figure 8-14, grey column]. The mean concentration measured for the pooled PCR products was 2.16 ng/μl with a standard deviation of 0.2 ng/μl. The whole volume of these solutions (31 μl) was then used for library preparation.

8.3.3.3 Evaluation of generated libraries

Libraries were prepared following the same experimental steps as when analysing the methylation controls using KAPA's library quantification kit. The average concentration (diluted 1:2000) was 178±30.8 nM, which is slightly lower than the one obtained in the first run (397±239 nM) but showing much less variation. There was only one library (blood sample 23) that resulted in a 3-fold decreased concentration probably due to experimental errors.

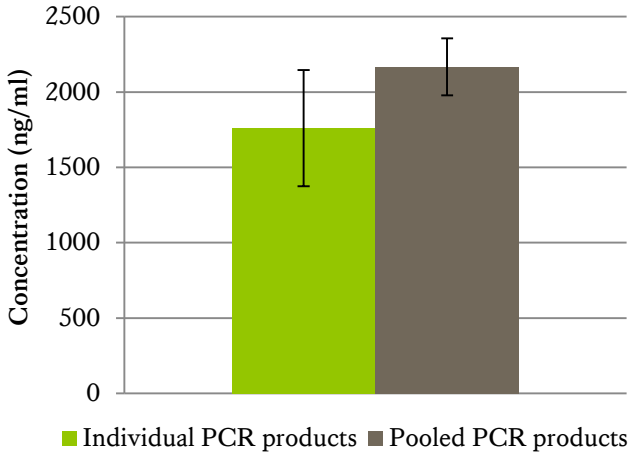


Figure 8-14. PCR product concentration (ng/ml) as measured by Qubit for both individual and pooled amplicons

The green column corresponds to the average concentration as measured in all individual PCR products for one blood sample (16). The grey column corresponds to the average concentration as measured in the pooled PCR products (2 µl of each individual product) of all blood samples (48). Errors bars represent standard deviation.

8.3.3.4 Distribution among samples and amplicons

As demonstrated in section 7.4.2.4, there was a large variation in read depths not only between samples but also among individual amplicons. It was believed that this could be due to differential PCR efficiency or variation during PCR product pooling and library preparation. In an attempt to minimise these differences, individual PCR products were quantified before pooling. However, it is generally accepted that Qubit does not only quantify PCR amplicons but also unused primers or bisulphite DNA strands. Taking into account all 34 samples, the average total reads per sample was 1,108,814±406,596 reads. This indicates a great coverage across all 16 markers; however, once again the distribution among them was uneven. Table 8-6 shows the minimum, maximum, average and range of reads for each CpG site. It was observed that the five markers showing the smallest average read number including cg17274064 (1,914), cg24450312 (2,055) and cg08370996 (3,730) correlated with previously obtained results [Figure 8-9]. Although these numbers might be considered small compared to other markers (e.g. cg04084157 – 154,988), based on confidence interval analysis Masser *et al* suggested that methylation quantification is not massively improved at sequencing read depths greater than 1,000 reads (Masser *et al.*, 2013). However, to increase the representation of these markers, their PCR assays could either be further optimised in the future or more PCR products could be added in the final pooled amplicons.

Table 8-6. Read numbers per individual CpG site (16)

CpG site	Read number			
	Minimum	Maximum	Average	St. Deviation
cg19761273	3,980	41,770	15,938	8,956
cg27544190	10,306	291,415	103,305	58,350
cg03286783	40,906	219,615	112,227	47,113
cg01511567	75,747	292,657	167,919	58,944
cg07158339	4,022	20,864	10,250	4,660
cg05442902	15,579	299,169	135,538	62,418
cg24450312	873	5,010	2,055	1,056
cg17274064	414	3,601	1,914	893
cg02085507	3,694	53,987	30,672	14,538
cg20692569	4,381	42,706	17,138	8,715
cg04528819	51,489	301,616	127,746	50,494
cg08370996	1,023	8,957	3,730	1,782
cg04084157	769	414,868	154,988	77,520
cg22736354	17,026	88,772	43,753	21,157
cg06493994	56,751	359,595	148,357	70,654
cg02479575	7,535	81,549	33,283	15,758

8.3.3.5 Reproducibility/Precision

In order to assess how precise methylation quantification is based on the proposed MiSeq® method, six technical replicates of the same blood sample were analysed. 200 µl of blood from a male aged 55 years were used to isolate genomic DNA, 100 ng of which were treated with sodium bisulphite in six replicates. The samples were then analysed as described in section 8.2.3 along with the rest of the blood samples. The obtained methylation data for all replicates are shown in Figure 8-15a. As illustrated from the generated graph, the variation in methylation quantification ranged from 0 to over 0.2 with an average of 0.061 [Figure 8-15b]. It should be noted that the standard deviation (SD) was generally low in most cases (<0.05) but three markers (cg07158339, cg05442902 and cg20692569) resulted in SD~0.1 while cg02085507 demonstrated the highest variation (0.22). Interestingly, cg02085507 showed one of the least accurate quantifications when the known methylation controls were analysed [Figure 8-11d]. In most cases, these observations correlate with the ones obtained by Masser et al (2013), where authors reported a maximum of 0.05 standard deviation when using their MiSeq®-based method.

(a)

CpG sites	Detected DNA methylation ratio for each replicate						Mean	St Dev
	1	2	3	4	5	6		
cg19761273	0.134	0.177	0.235	0176	0.195	0.166	**0.181**	**0.033**
cg27544190	0.087	0.044	0.041	0.085	0.061	0.036	**0.059**	**0.023**
cg03286783	0.002	0.002	0.002	0.002	0.002	0.002	**0.002**	**0**
cg01511567	0.102	0.116	0.058	0.070	0.128	0.087	**0.093**	**0.027**
cg07158339	0.334	0.569	0.458	0.338	0.305	0.353	**0.393**	**0.101**
cg05442902	0.365	0.138	0.229	0.292	0.179	0.094	**0.216**	**0.100**
cg24450312	0	0.001	0.001	0.001	0.001	0	**0.001**	**0**
cg17274064	0.121	0.049	0.167	0.134	0.139	0.066	**0.113**	**0.046**
cg02085507	0.303	0.532	0.410	0.772	0.887	0.572	**0.579**	**0.219**
cg20692569	0.145	0.349	0.368	0.255	0.077	0.169	**0.227**	**0.117**
cg04528819	0.120	0.060	0.085	0.132	0.075	0.078	**0.092**	**0.028**
cg08370996	0.129	0.093	0.006	0.010	0.205	0.007	**0.075**	**0.082**
cg04084157	0.013	0.156	0.002	0.006	0.055	0.003	**0.039**	**0.061**
cg22736354	0.055	0.142	0.056	0.037	0.052	0.113	**0.076**	**0.041**
cg06493994	0.095	0.219	0.056	0.092	0.144	0.092	**0.116**	**0.058**
cg02479575	0.039	0.002	0.001	0.051	0.092	0.093	**0.046**	**0.041**

(b)

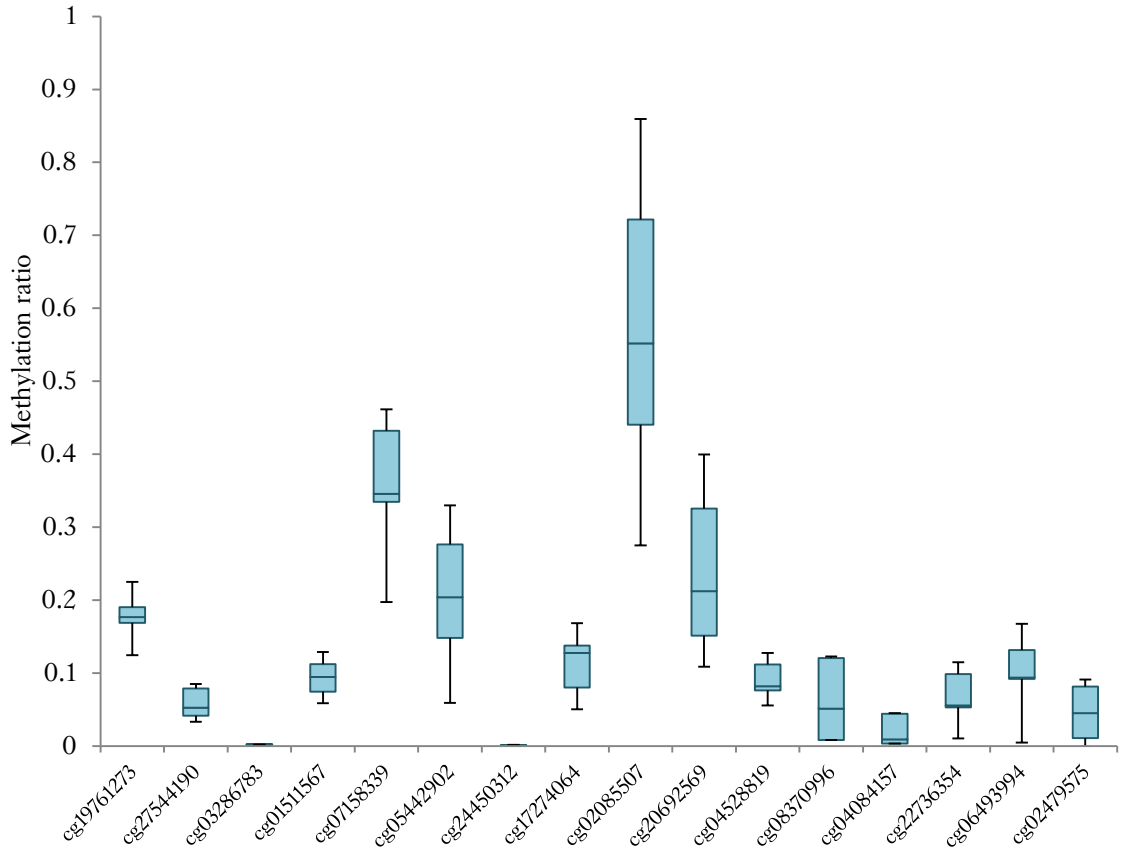


Figure 8-15. Observed methylation levels among six technical replicates of a blood sample
(a) Reported methylation ratio of all replicates together with calculated mean and standard deviation for each of the 16 CpG sites. As shown, most CpGs had <0.05 standard deviation; however there were seven markers (highlighted in red) that demonstrated a higher variation. (b) Box-and-whisker plots for each individual CpG sites indicating the minimum, maximum, median and interquartile range of detected methylation.

8.3.3.6 Age prediction accuracy

To assess the capability of the proposed method to predict age, blood samples' methylation values for all 16 age-associated CpG sites were inserted into the previously described age prediction ANN model [Figure 7-13]. One sample (no. 23) resulted in reads <1,000 for most markers so it was excluded from analysis; therefore, a total of 33 blood samples were applied. Taking into account the outcomes of the linearity graphs and employing the individual equations described in Table 8-5, methylation values were corrected as described in section 2.4.1.5. Lastly, for predictions both original and 'corrected' methylation values were used. Figure 8-16 presents the predicted *vs*. true age graph as well as the observed predicted error for all individuals using both original (purple) and corrected data (green). As a whole, the age correlation using the original data was 0.67 with a mean absolute error of 13.32 years while the age correlation using the corrected data was 0.66 with a mean absolute error of 10.65 years.

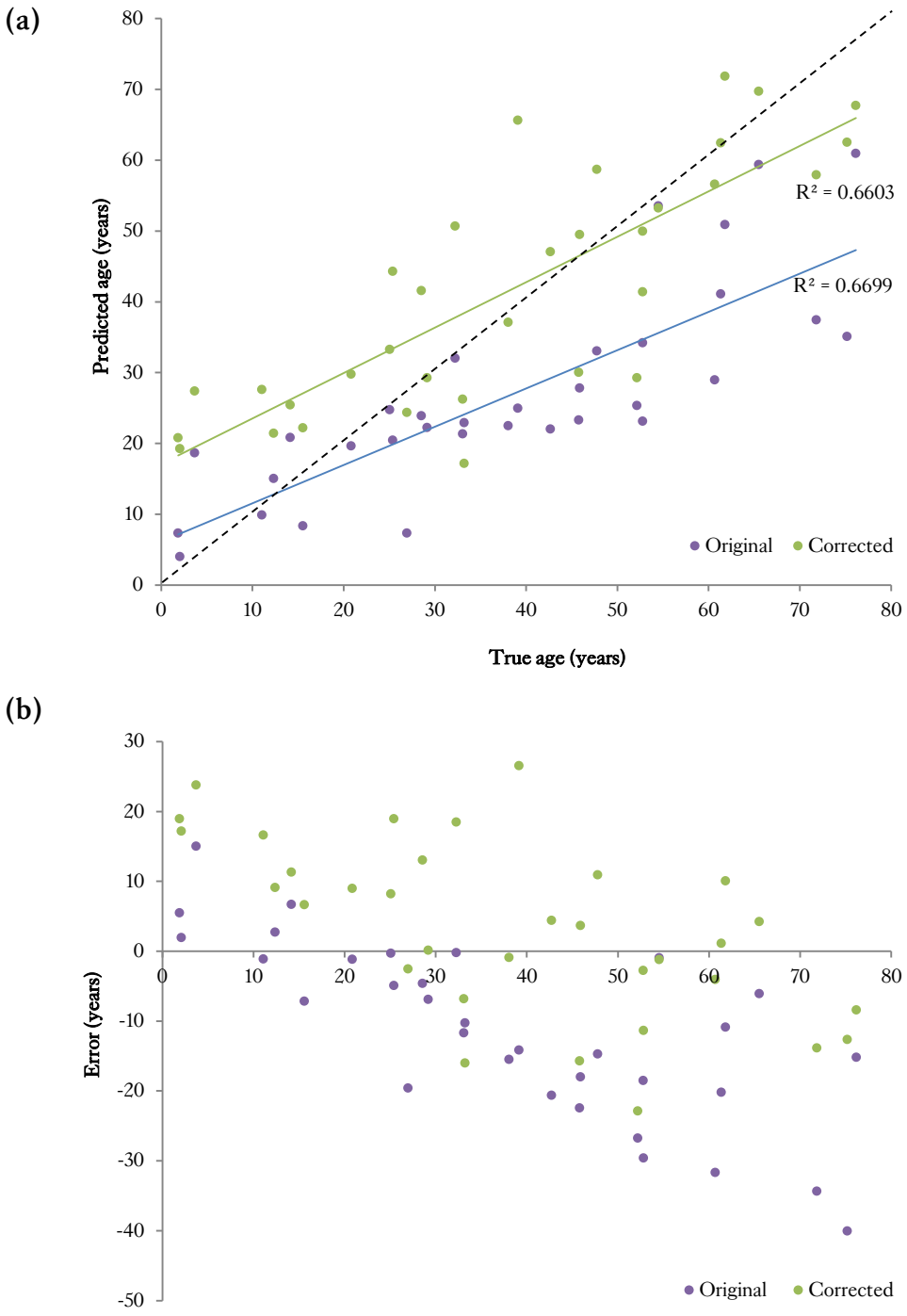


Figure 8-16. Age predictions using the developed NGS method (n=33)

(a) Predicted vs. true age for both sets of methylation data (original and corrected); the coloured lines correspond to linear correlation (R^2=0.67 and R^2=0.66 respectively) while the black dotted line represents the 'ideal' age correlation, (b) predicted error in years for both sets of methylation data (original and corrected).

8.4 Final Remarks

As shown in Figure 7-11, the observed age-associated DNA methylation changes of the selected 16 CpG sites were quite small, especially for certain markers. It was also shown that there were issues with the precision and accuracy of bisulphite Pyrosequencing® when trying to predict age; therefore, it was decided that the potential of next generation sequencing is explored since it was previously reported to be able to accurately measure methylation levels (Masser *et al.*, 2013). In this study, an NGS protocol based on MiSeq® specifically developed by Illumina for forensic applications was developed and validated through linearity analysis and the use of pre-defined DNA methylation standards. Issues regarding sample representation and marker distribution were observed, but it is believed that further optimisation will potentially reduce these problems. The proposed method showed that is capable of accurately quantifying methylation levels; however by testing two sets of standards amplified by either 30 or 45 cycles revealed that PCR amplification biases occurred. Therefore, methylation correction was performed to assess if accuracy is improved when using the obtained methylation values for age prediction. The method's capability on predicting age was tested in a total of 33 samples, where the mean absolute error was 13.32 years (original data) and 10.65 years (corrected data). Prediction differences within age groups were observed, with prediction of older people being significantly improved when correct values were used. In conclusion, although the proposed method resulted in poorer predictions comparing to the developed ANN model (Chapter 7), it seems to be very promising for age prediction; however, future research on optimising the protocol is needed to improve accuracy.

9 Final discussion and future work

9.1 Tissue-specific mRNA profiling

For the purpose of identifying the tissue source of a recovered body fluid stain, forensic researchers have focused on identifying suitable mRNA molecules showing tissue-specific expression (Fleming & Harbison, 2010a; Haas *et al.*, 2009a). In this study, a number of multiplex PCR systems developed by other research groups were assessed as part of ongoing European collaborative projects to assess inter-laboratory variation in the performance of the developed methods, prior to applying the proposed markers in forensic casework. Overall, both tissue-specific (EDNAP) and multi-tissue (EuroForGen) mRNA-based methods were found to be very sensitive and able to accurately differentiate between forensically relevant body fluids and tissues from minute amounts of starting material. However, differences in marker specificity were observed when testing various mock casework samples. Although false negative results were regularly obtained in very low quantity samples, false positive results as a result of a low expression of tissue-specific mRNAs in non-target tissues were also seen. It is believed that the use of endogenous controls (housekeeping genes) should not only be used for the confirmation of the presence of cellular material (as in TissueID 20plex) but also in a more quantitative way to potentially correct for these 'stochastic' variations, more likely caused by the large and varying amount of cells analysed (sample overload). It was observed however, that by using replicates of the same sample and introducing a category of 'sporadically observed' in the interpretation guidelines, false positives could be eliminated.

Although for body fluids like semen and blood these problems rarely occured, the identification of vaginal secretion using mRNA profiling proved to be more challenging due to the occasional detection of menstrual blood-, saliva- or skin-specific markers [Table 4-5 & 4-6]. It is understood that menstrual blood- and vaginal fluid-specific mRNAs can be affected by the menstrual cycle as well as by individual differences in menstrual flow; however, to the best of my knowledge this has not yet experimentally been tested. In the future, gathering information regarding the day of menstrual cycle, possible use of contraceptives or other similar information when collecting vaginal fluid and menstrual blood samples could help with identifying the source of these variations. In addition, it is believed that due to the epithelial tissue of origin, vaginal secretion, saliva and skin might display similar gene expression patterns.

This was highlighted by the detection of putative vaginal fluid-specific markers in buccal and saliva stains as well as the detection of skin markers in both saliva and vaginal fluid using the tissue-specific EDNAP multiplexes [Table 4-5]. Nevertheless, the use of bacterial-specific markers showed themselves to be very promising for vaginal fluid detection since it is known that the presence and concentration of particular *Lactobacillus* species are specific within this body fluid. However, research is still needed to select the best bacterial species since, for example, one marker (Ljen) was found to be co-expressed in urine, buccal, skin, saliva and menstrual blood samples [Table 4-5].

Although it is generally accepted that mRNA profiling could serve as a suitable confirmatory test for differentiation among tissues, future optimisation of existing multiplex mRNA-based methods would allow for a more accurate identification. It is believed that a test that would simultaneously analyse markers for various tissues in a single reaction would surpass methods that are tissue-specific (determining only one tissue at a time). Optimising the proposed 20plex system (EuroForGen) could result in a better performance in casework. Adjustments could include the addition of more specific markers gathered by the analysis of tissue-specific assays (EDNAP) and the removal of mRNAs that seem to give non-specific signals (such as the skin marker LOR). Also, further method optimisation will ensure equal amplification efficiencies across mRNA molecules since in this study a large variation in detected peak heights (rfu) was observed. Lastly, from the analysis of complex mock casework containing up to three different tissues in a single stain, as well as from the data interpretation exercise, it was recognised that guidelines on interpreting RNA profiles were needed and forensic scientists would require special training in interpretation.

Careful interpretation is especially necessary when menstrual blood markers are interpreted. When analysing a menstrual blood/whole blood mixed sample (section 4.3.4.3), although the identification of menstrual markers was successful, the detection of blood was masked under the 'observed and fits' category due to the co-expression of markers in peripheral and menstrual blood. It was suggested that in these cases the peak height of blood peaks as well as the obtained DNA profiles could be used to assess if these are due to the 'true' presence of non-menstrual blood or due to co-expression in menstrual blood. However, Harteveld *et al* have shown that not only

RNA peaks cannot be associated to DNA peaks but also RNA peak heights can vary between replicates of the same sample (also observed in this study) (Harteveld *et al.*, 2013).

Except for further optimisation of the TissueID 20plex assay for tissue identification, future work could include the analysis of a larger set of mock casework samples, where lessons can be learned regarding the best experimental and interpretation strategy. Furthermore, following the example of other countries such as the Netherlands and New Zealand, the application of mRNA profiling for tissue identification should be routinely used, therefore efforts towards setting up validation and interpretation criteria is vital.

9.2 Tissue-specific DNA methylation profiling

Although forensic researchers have investigated in depth the potential of applying mRNA profiling in tissue identification (Gomes *et al.*, 2013; Lindenbergh *et al.*, 2013) and research on developing robust assays is still ongoing (Xu *et al.*, 2014), it is thought that in certain forensic scenarios its application would not be feasible. For example, RNA is not yet routinely analysed in forensic laboratories, and also, for cold cases it is usually only the original stains or extracted DNA samples that have been stored (and are often several years old). Considering the instability of RNA molecules (even if studies have shown successful detection of mRNA profiles in aged stains (Kohlmeier & Schneider, 2012)), a method based on DNA would be preferable.

For this purpose, DNA methylation patterns among tissues were evaluated for their potential in differentiating tissues. When the present study was initiated, no research had been published assessing the application of epigenetic markers in forensic tissue identification, but the potential as well as challenges of such approach were soon apparent (Frumkin *et al.*, 2011; Lee *et al.*, 2012a). As an example, even though Frumkin *et al* proposed the use of seven loci for the simultaneous identification of blood, semen, saliva, skin, urine, menstrual blood and vaginal secretion, Gomes *et al* failed to reproduce their results on skin identification (Gomes *et al.*, 2011).

It is known that DNA methylation regulates gene expression and controls cellular differentiation during development (Shen *et al.*, 2007; Song *et al.*, 2009); however, in the medical field differential epigenetic patterns among tissues are mainly assessed in

order to evaluate if observed differential patterns in diseased tissues are meaningful events or part of natural tissue-to-tissue variation. The main challenge faced in this study regarding the identification of suitable CpG sites was the lack of methylation data in forensically relevant body fluids such as vaginal fluid and menstrual blood. Therefore, an extensive validation of the selected markers as well as the identification of novel ones was needed to assess their specificity.

Firstly, it was thought that since DNA methylation is involved in gene silencing (Newell-Price *et al.*, 2000), it would be interesting to investigate the methylation status of genes included in the proposed mRNA-based systems. For example, it has been previously reported that the blood-specific haemoglobin genes are regulated by DNA methylation (Bartzeliotou & Dimitriadis, 1989), while evidence is also available regarding the DNA methylation of semen-specific protamine genes and its involvement in determining their expression during spermatogenesis in mouse (Trasler *et al.*, 1990). Furthermore, it has been shown that for MUC4, previously suggested as a vaginal marker (Juusola & Ballantyne, 2005), DNA methylation in its 5'-flanking region plays an important role for its gene expression in carcinomas (Yamada *et al.*, 2009). However, the exact chromosomal location of potentially useful CpG sites was not always known, and therefore a promoter screening approach was decided to be more suitable. For this purpose, the promoter regions of selected genes such as HBB and MUC4 were used for the design of Sanger sequencing assays since this method allows for the sequencing of up to 500-600 bp in a single reaction. However, it was soon evident that amplifying long fragments of bisulphite-treated DNA is very challenging due to its low complexity; and primer re-design or extensive optimisation did not necessarily eliminate non-specific products present that interfered with methylation detection (data not shown).

Therefore, the selection of previously reported differentially methylated CpG sites as well as the development of a method that would allow for the amplification and sequencing of shorter fragments were needed. The capabilities of bisulphite Pyrosequencing® on quantifying methylation levels have been previously assessed (Dejeux *et al.*, 2009) and its advantages over other methods have also been established (Reed *et al.*, 2010). Therefore, using a commercially available CpG assay for HBA1, a Pyrosequencing® method was validated for its sensitivity, bisulphite conversion

efficiency as well as reproducibility. Even though the results were very promising, the four CpG sites quantified in this assay failed to demonstrate differential methylation patterns among blood, semen and saliva, results that further support the need for either identifying existing tissue-specific CpG sites or discovering novel ones through a genome-wide analysis.

In a first approach, a reported blood-specific region of the EFS gene consisting of ten CpG sites was investigated using various forensically relevant tissues and, as shown in Figure 5-9, its potential for blood detection was confirmed. However, saliva demonstrated a wide range of average methylation levels among individuals even though buccal samples showed a lower one. It is believed that this could be due to the presence of leukocytes in saliva, due to either immunological conditions or simply due to gum bleeding. Nevertheless, one of the analysed CpG sites (CpG 4) was found to be the most useful marker since it was only highly methylated (>0.8) in blood, allowing not only for the discrimination between blood and saliva but also between blood and menstrual blood. Additionally, the assay successfully passed the validation phase showing that methylation levels among a total of 65 blood samples were consistent and reproducible [Figure 5-11 & 5-12]; this was also demonstrated in aged and degraded samples (e.g. the expected methylation profiles were successfully obtained from stains deposited 18 years prior to analysis).

In a second approach, analysis of genome-wide methylation data gathered from blood, semen and buccal cell samples identified a set of eleven body fluid-specific CpG sites. In this case, following analysis with bisulphite Pyrosequencing® only the selected semen markers were confirmed as suitable. This was once again due to similar methylation profiles among blood and saliva samples, which further supports the previous theory on the presence of naturally occuring white blood cells in saliva (Calonius, 1958), further supported by the difference in methylation between buccal cells and saliva as shown in Figure 5-16b and 5-17b. Nevertheless, one CpG site (cg13763232, BLM1) demonstrated a similar pattern with EFS, showing high methylation levels only in whole blood (>0.85). Interestingly, this CpG site belongs to the promoter region of the solute-carrier family 6, member 6 (*SLC6A6*) gene that encodes for a sodium- and chloride-dependent transporter of the neurotransmitters taurine and beta-alanine (Han & Chesney, 2003). Taurine plays an important role in

many biological activities including osmoregulation, membrane stabilisation and anti-oxidation (Han *et al.*, 2006). It has been reported that this gene is regulated by controlling transcription factor binding sites in its promoter; therefore, the involvement of DNA methylation in this control cannot be excluded. Lastly, a very recent study implicated its role in colorectal cancer via inactivation of cell apoptosis (Yasunaga & Matsumura, 2014).

As shown in similar studies (An *et al.*, 2013; Lee *et al.*, 2012a; Madi *et al.*, 2012), identifying semen-specific methylation patterns has shown to be rather simple. This comes as no surprise since sperm cells have a unique composition in terms of histones and proteins. This was highlighted further for DNA methylation, since initial selection of highly differentially methylated sites between semen and blood/buccal cells using the genome-wide methylation data obtained by Rakyan *et al* (2008 & 2010) resulted in hundreds of potential markers. As shown in Figure 5-17, all four tested CpG sites demonstrated semen-specific differential methylation; however, two of them (cg01318557 and cg05656364) did not seem to be sufficiently robust.

Looking at the biological functions of the involved genes for the best two semen-specific CpG sites to potentially understand and justify the observed differential methylation in semen, some very interesting observations were made. The sequence analysed by the SEU1 assay belongs to a gene encoding for the solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 31 protein (SLC25A31) which catalyses the exchange of cytoplasmic ADP with mitochondrial ATP across the mitochondrial inner membrane. It is believed that it mediates energy generating and energy consuming processes in the distal flagellum, possibly as a nucleotide shuttle between flagellar glycolysis, protein phosphorylation and mechanisms of motility (Dolce *et al.*, 2005). It has previously been reported that its SLC25A31 mRNA transcripts are exclusively present in liver, testis and brain (Dolce *et al.*, 2005); therefore, differential DNA methylation in its promoter (where cg04382920 is found) could regulate its expression. A recent study supports this assumption, as SLC25A31 was found to be one of the few differentially methylated genes in semen following a genome-wide analysis of 38 semen samples (Schutte *et al.*, 2013). It was reported to be involved in spermatogenesis and associated with inflammation and autoimmune processes interfering with fertility.

Moreover, cg11768416 (SEU2) belongs to the gene encoding for the coiled coil glutamate rich protein 1 (Ccer1 or C12orf12). Remarkably, this gene has been previously reported to have dense promoter CpG island methylation and gene silencing in normal tissues except testis and sperm (Shen *et al.*, 2007); a finding that was confirmed in this study. Authors tested various healthy and diseased tissues and found that hypomethylation of this gene in non-semen tissues is associated with gene activation in cancer.

Furthermore, using the proposed markers, successful detection in aged stains of up to 16 years old as well as high sensitivity (using as little as 50 pg of starting DNA material) was obtained making their application in forensic casework possible. The sensitivity of DNA methylation assays for tissue identification have also been demonstrated in the study by Larue *et al* (2013), where positive results for semen were obtained down to 31 pg using the proposed DSI-Semen™ kit developed by Wasserstrom and co-workers (2013).

Lastly, a set of six immune-cell specific methylation markers was tested using all types of forensically relevant tissues mentioned so far, together with nasal fluid and nasal blood. Although semen-specific DNA methylation profiles of two loci (AMP1730 and AMP2007) were confirmed and one potential blood-specific marker was identified, the rest of the assays failed to demonstrate differential methylation levels for the other body fluids as originally suggested by our European collaborator (personal communication) [Figure 5-3].

Interestingly, methylation results obtained by the analysis of nasal blood showed that the differentiation between different 'types' of blood - whole, menstrual and nasal - is particularly challenging. However, it is noteworthy to mention that as expected, nasal blood demonstrated a 'mixed' profile between nasal fluid and blood, which could potentially be used for their differentiation if samples of the individual in question are available. It should also be noticed that a few urine samples gave positive reactions when testing the semen-specific markers. However, 74% of these urine samples belong to male volunteers and these false-positive results could be explained since both male urinary and reproductive systems share the same tract.

As a comment, it should be noted that caution is needed when applying DNA methylation markers for tissue identification since changes in DNA methylation patterns have been reported in various diseases (Teschendorff *et al.*, 2012; Tsai & Baylin, 2011). With regards to semen detection it has been shown that alternations in sperm DNA methylation patterns in particular loci are seen with low sperm motility and different types of male infertility (El Hajj *et al.*, 2011; Hammoud *et al.*, 2010; Pacheco *et al.*, 2011). In fact, in this study there were two semen samples in particular that showed an 'unexpected' methylation ratio for most semen-specific assays tested as shown in Figure 5-17b for SEM2, SEU1 and SEU2, in Figure 5-22 for AMP1730 and in Figure 5-23 for AMP2007. Interestingly, these two samples resulted in a 10-fold decrease in DNA yield following DNA isolation using the same starting material as the rest of the semen samples, potentially indicating a smaller number of spermatozoa present.

The observations above, including the implication of some of these markers in cancer via an 'abnormal' methylation profile as well as the possible 'connection' between low sperm counts and altered methylation levels, highlights the need for validation of the proposed markers prior to implementation by analysing diseased samples. Ideally, together with samples from 'healthy' volunteers, samples from patients suffering from various diseases should be co-analysed so that a better representation of the general population is achieved. Due to the ethic restrictions, gathering personal data regarding disease status from crime scene stains is not permitted; therefore, all possible reasons that could lead to altered methylation profiles should be taken into account.

As a general comment and following this study, it can be concluded that an important drawback of applying DNA methylation profiling into forensic casework is the analysis of mixed stains. Since DNA methylation results are presented in a more quantitative manner compared to the mRNA profiling (percentages rather than the presence/absence of peaks), a result regarding one tissue-specific marker would be insufficient to report the tissue source. Therefore, a serial analysis including at least two or three CpG sites per tissue (similar to the proposed mRNA strategy) would be necessary in order to include or exclude the presence of a body fluid. Afterwards, the obtained ratios for each marker could be used to figure out which tissues are involved in the case of mixed stains.

Future work regarding the tissue-specific DNA methylation profiling could include not only an extensive validation of the blood- and semen-specific markers found in this study, but also the investigation of more genomic loci showing potential tissue-specific DNA methylation patterns. Using the same approaches but further developed (for example, collecting genome-wide methylation data for all forensically relevant tissues) could reveal more markers of interest that could be selected during the discovery phase for assay development. Also, to account for potential, common inter-individual differences in DNA methylation and consequently, gene expression, a larger dataset of body fluids and tissues should be collected.

9.3 Implementation of tissue identification to live casework

As gathered from both the literature and this study, DNA methylation markers can be a useful additional method for use in forensic casework. As demonstrated during the validation phase of the proposed tissue-specific CpG sites, these seemed to be stable not only over time but also under harsh conditions (high temperatures, outdoors, UV-degraded), highlighting their potential for application in forensic-type samples. Any newly proposed tissue-specific marker need to undergo an externsive validation phase, which following this study can be further extended. To our knowledge, mRNA and DNA methylation profiling have not yet been implemented in forensic casework in the UK and following this study, we were able for the first time to apply the best tissue-specific markers for confirmatory tissue identification in two cold cases (data not shown).

The applicability of the two semen markers identified through genome-wide methylation data analysis (SEU1 and SEU2) was tested in a sexual assault cold case where differentiation of semen *vs* vaginal fluid was required. The advantages of DNA methylation profiling in comparison to mRNA-based approach were obvious since there was only a DNA sample available. As a result, we were able to conclude that it was very unlikely that semen was present in the analysed DNA sample since the observed methylation levels at both loci were 0.85 and 0.88 (semen exhibits low levels of methylation for both markers). A level of uncertainty due to potential natural inter-individual variations could be eliminated in the future by analysing more semen samples as part of the assessment of markers' specificity as well as by introducing guidelines on interpreting DNA methylation profiles for forensic casework. Finally,

having more information regarding the sperm count could be useful, however this is often not possible when only the DNA sample rather than a semen stain is retained.

Additionally, the usefulness of combining mRNA and DNA methylation profiling was assessed in a case involving a reported sexual assault of a female in menses. In this case there was only a stained duvet cover available. The stains on this item had returned positive for blood using presumptive testing; however, scientists were unable to say if the stains were of peripheral or menstrual origin. Using the TissueID mRNA 20plex, validated in Chapter 4, together with the blood markers proposed in Chapter 5 (EFS and AMP1404), we were able to conclude that the stain of interest was most likely menstrual blood. Firstly, this was achieved by the detection of peaks for menstrual secretion (MMP10), blood (HBB) and housekeeping genes (ACTB and 18S rRNA) using the TissueID 20plex. The presence of the blood peak would be expected in a menstrual blood stain, since it is known that co-expression of blood and vaginal fluid markers is common in menstrual blood (Lindenbergh *et al.*, 2012). Secondly, detected EFS methylation was very low (<0.2) for all CpG sites, a profile that contradicts the one of whole blood, which is highly methylated [Figure 5-9]. Similarly, the AMP1404 locus was found to be completely methylated (>0.95), whereas in this study, whole blood samples resulted in partial methylation (average of 0.7) [Figure 5-22a].

This case particularly demonstrates the advantages of using a combined genetic and epigenetic approach, since the presence of menstrual blood was confirmed by means of menstrual blood-specific mRNA profiling and the presence of whole blood was excluded by using blood-specific CpG sites (since menstrual blood-specific CpG sites are not available at the moment). Together with the results obtained from presumptive testing, conclusions drawn from mRNA- and DNA methylation-based tests could be further strengthened.

9.4 Age-associated DNA methylation profiling

The ability to accurately estimate a person's chronological age would be a great advantage in police investigations as it could provide significant investigative leads. Although there have been various approaches to estimate age in human remains and living individuals, current methods suffer from low accuracy of age prediction, usually producing an estimate with a large age range. Since DNA methylation is known to act

as an interphace between the genome and environment, age-associated DNA methylation patterns have been observed (Gentilini *et al.*, 2013; Heyn *et al.*, 2012). Also, using genome-wide methylation analysis, researchers have been able to predict chronological age in a variety of tissues (Horvath, 2013); however, they all suffer from low sensitivity and use hundreds of loci which make them not applicable in forensic casework in their current form. In this study, a total of 55 potentially age-associated CpG sites were investigated. These CpG sites have been reported in two of the largest and highest-resolution genome-wide methylation studies in ageing (Hannum *et al.*, 2013; Horvath, 2013).

Regarding the first approach, a set of ten CpG sites was selected for further investigation. Most of the genes associated with the proposed markers were found to be linked with age-related conditions, such as cancer, arthritis and Parkinson disease. In particular, two of the selected markers belong to the transcription factor KLF14 (Krüppel-like factor 14), which is known as the 'master regulator' of obesity and other metabolic traits (Small *et al.*, 2011). Furthermore, thanatos-associated protein 7 (THAP7, 'thanatos' meaning death in greek) is also another protein involved in transcriptional regulation (Macfarlan *et al.*, 2006). Also, two markers included in the initial set that are in close proximity, lie within the gene CD248, which encodes for a transmembrane glycoprotein found on the surface of activated perivascular and fibroblast-like cells and has a known function in inflammatory arthritis (Maia *et al.*, 2010). Lastly, mutations in a transmembrane lysosomal P5-type ATPase gene (ATP13A2) have been reported to unravel an essential role in lysosomal function and cell viability and suggested as a therapeutic target against Parkinson's disease (Dehay *et al.*, 2012).

Subsequently, their epigenetic drift in blood of individuals belonging to different age groups was assessed and, as shown in Figure 7-5, it was changing slightly over time but regression analysis revealed no statistically significant correlation ($p>0.05$). Since these markers have been previously reported to be age-associated in the literature (Hannum *et al.*, 2013), it is believed that this could be due to various reasons. It could be due to the small sample set or it could be due to failure of bisulphite Pyrosequencing® to accurately detect such small methylation differences (<0.1). An average of 5% standard deviation has been obtained in other bisulphite Pyrosequencing® assays

(section 5.3.1.3 and Figure 5-12) but there were cases where higher deviation was also observed. Moreover, in this method, a total of 45 PCR cycles was used which could introduce amplification bias and alter 'true' methylation levels. These results were confirmed by multiple regression analysis which revealed that only two markers from this set (cg05442902 and cg20426994) were important for age prediction using this particular dataset.

To assess potential non-linear correlation of these markers with age, artificial neural networks (ANN) were applied. ANN models have been successfully applied in medical research and seemed to be able to detect complex relationships in various biological functions (Patel & Goyal, 2007). Initially, using a methylation dataset of 65 samples, ANN analysis resulted in an age prediction model with a correlation between observed and predicted age of 95% and an average error of 4.27 years. However, the sample set used was considered relatively small (six samples in the verification set and only three in the blind test) so analysing more samples was required before drawing any conclusion. Surprisingly, the analysis of a larger dataset failed to reproduce the above age prediction accuracy, indicating that a different approach should be followed to both overcome the method's low accuracy and reproducibility but also to take advantage of reported age-related markers with stronger association with age.

Thus, a set of 45 CpG sites was selected from a genome wide methylation study that used not only blood samples but also other tissues, potentially resulting in more robust results. Although in this study, only 45 CpG sites were selected from the Horvath study (2013) which included a total of 353 age-associated markers, a future approach could include all these markers together with newly published potential markers. Thus, a much larger pool of age-associated CpG sites (up to thousands) could be created and used for age prediction analysis.

Once again, all these 45 sites belong to genes involved in age-related conditions. For example, cg19761273 is associated with casein kinase 1 (CSNK1D), which is a serine-threonine protein kinase involved in essential cell pathways including circadian rhythms and DNA repair. It is believed that CSNK1D has a role in arranging the microtubule network during mitosis to prevent DNA damage (Behrend *et al.*, 2000). On the same theme, the structure specific recognition protein 1 (SSRP1) linked with cg01511567 seems to be crucial to anticancer mechanisms since it forms a chromatin-

specific transcriptional elongation factor (FACT) that interacts specifically with histones and prevents DNA damage (Yarnell *et al.*, 2001). Additionally, cg03286783 belongs to the cancer susceptibility candidate 4 gene (*CASC4*), which appears to produce a single-pass membrane protein related to Golgi membrane protein 1 (GOLM1); increased expression levels of CASC4 have been associated with proto-oncogene overexpression in 30% of breast and 20% of ovarian cancers (Oh *et al.*, 1999).

Moreover, cg05442902 is not only associated with THAP7 as shown earlier, but also with the purinergic receptor P2X-like 1 (P2RXL1) known for its involvement in a broad range of physiological processes via extracellular adenine nucleotides. These include inflammatory and immune responses and neurotransmitter release that could be affected by ageing (Kendall Harden *et al.*, 1995). A recent study has also linked the thyroid hormone receptor interactor 10 (TRIP10) (cg02085507) with the regulation of cancer cell growth and death in a cancer type-specific manner; in fact differential DNA methylation of its gene has been suggested to promote cell survival or death (Hsu *et al.*, 2011). Once again, one of the markers belongs to the transcription factor KLF14 (cg04528819) which, as mentioned earlier, is known as the 'master regulator' of obesity and other metabolic traits (Small *et al.*, 2011). Lastly, the nerve growth factor VGF associated with cg04084157 has been linked with the age-associated Alzheimer's disease, where its expression levels were found to be altered (Carrette *et al.*, 2003).

Analysis of 1,156 blood samples gathered from a total of 7 genome-wide methylation studies using both multiple and stepwise regression as well as ANN analysis revealed that an accurate prediction could be obtained by using 16 out of the 45 CpG sites. Multivariate regression revealed that none of the available factor variables (gender, ethnicity) influenced age-associated DNA methylation patterns in a statistically significant way. Although multiple linear regression analysis resulted in an accurate age prediction model (mean average error of 4.89 years), it was shown that once again the application of ANN analysis resulted in a more accurate model. In particular, the ANN age prediction model seemed very promising as it resulted in a correlation between predicted and true age of 96% and a mean absolute error of 3.3 years (standard deviation=3.7 years) [Figure 7-13]. It is believed that ANN models have the

ability to recognise complex patterns, which are often observed in complex traits like chronological age.

Furthermore, analysis of 1,011 diseased blood samples (of various pathological conditions) resulted in a less accurate prediction, which was expected. Even though the prediction for individuals suffering from conditions like schizophrenia (error=5.03 years) did not seem to greatly change, the samples from blood disorders resulted in a poor age prediction (error=12.74 years). Even though it is believed that the small size of the sample dataset could have an influence in these observations (n=105), it is suggested that CpG sites involved in blood diseases such as anaemia and bone marrow disorders should be excluded in future work so that these variations are avoided. One has to bear in mind that DNA methylation-based age prediction models actually predicts the biological age of tissues, rather than chronological age; therefore, 'stressed' cells could show an accelerated aging effect. In fact, as shown in Figure 7-14, the age prediction in diseased samples was getting worse in older individuals (>60 years old).

Also, the influence of environment could be demonstrated by analysing 53 pairs of monozygotic twins. In general, the obtained age prediction resulted in an average mean absolute error of 7.07 years; however, it was noted that the model gave either very accurate or very inaccurate predictions. The age prediction differences within the twin pairs were 2.65 years on average and were not statistically significant as obtained by paired t-test analysis (p=0.99), suggesting that age acceleration could be inherited. These findings further support the hypothesis that environmental influences including lifestyle, smoking, diet or disease susceptibility could influence DNA methylation patterns; and as long as this information is not available, age prediction might still remain poor.

Lastly, the possibility of applying the proposed model built on blood methylation profiles on other healthy tissues was assessed by analysing four different tissues including saliva, cervix, skin and muscle. Initially, by testing 15-28 samples per tissue, the ANN model failed to produce an accurate prediction. It was concluded that this could be not only due to the different biological nature of these tissues but also because of the small size of sample set used. It was therefore decided to focus only on saliva and cervix as there were few methylation datasets available for the other two tissues. The ANN was retrained using a total of 265 saliva and 167 cervix samples and as shown in

Figure 6-22 the potential of applying the selected blood CpG sites in other tissues seemed promising. However, it is believed that age prediction would be improved if more samples are gathered. Also, it is recommended that the potential of other CpG sites from the pool of 353 CpG sites suggested by Horvath is explored so that possible tissue-specific age-associated CpG sites can be investigated.

Considering that a similar prediction accuracy was observed when using 353 CpG sites in Horvath's study (age correlation of 0.96 with a median absolute error of 3.6 years), predicting age using a smaller number of CpG sites could be possible. This is also supported by a study where researchers obtained a mean absolute deviation (MAD) from chronological age of 5.4 years using only three CpGs in blood (Weidner *et al.*, 2014). Furthermore, in the forensic field, two very recent studies were published further supporting the possibility of predicting age using only a small set of DNA methylation markers. Firstly, applying a very different methodological approach (methylation-sensitive representational difference analysis, MS-RDA), Yi *et al* identified eight age-associated gene fragments. Using 65 samples they built a regression model by using two CpG sites from each fragment (16 in total) that explained 95% of the variance in age (Yi *et al.*, 2014). Also, even though the correlation with age was good (r=0.91), the small number of samples (n=65) as well as the large amount of DNA used (3 μg) makes its forensic application still questionable.

Secondly, a recent study examined the methylation status of seven CpG sites located in the promoter region of the previously reported age marker ELOVL2 by means of bisulphite Pyrosequencing® (Zbieć-Piekarska *et al.*, 2015). A total of 303 blood samples from individuals aged 2-75 years old were tested and multivariate regression analysis allowed for the selection of the best two CpG sites (namely, CpG 5 and CpG 7). Notably, CpG 7 alone explained 83% of variation in age. The developed model (R^2=0.859) explained 86% of variation in age and gave an average error of 6.85 years (MAD=5.03 years). Following validation with an additional set of 124 samples, 68.5% were predicted with an error of ±7 years; interestingly individuals above 60 years old resulted in poorer predictions (Zbieć-Piekarska *et al.*, 2015). Moreover, sensitivity analysis revealed that the same prediction rates were obtained with as little as 2.5 ng of starting DNA material while the storage of bloodstains at room temperature (up to 15 years) did not seem to correlate with prediction success rate either. This study seems

to be very promising for forensic age prediction which could potentially be improved if other age-associated markers apart from ELOVL2 are incorporated into the model.

As shown in the present study and considering the weak age effects on individual CpGs, the contribution of each marker was found to be different with cg22736354, cg19761273 and cg02479575 showing the strongest input while cg03286783 demonstrated the weakest influence in the model mirroring the methylation graphs on Figure 7-11. For a future approach it is recommended that the selection of age-associated CpG sites is not only based on age coefficient values but also on their methylation change with age. Potentially, the age correlation of all 353 CpG sites proposed in Horvath's study could be assessed via plotting their methylation levels against the subjects' age and the CpGs showing a more noticeable variation could be chosen. Consequently, CpG sites with low contribution in the proposed model could be replaced by others with greater potential in age prediction as reported in the studies above (Yi *et al.*, 2014; Zbieć-Piekarska *et al.*, 2015).

Since this is the first study where artificial neural networks are implemented in age prediction using DNA methylation markers, it can be concluded that ANN could offer a useful alternative to the regression models reported in the literature as it significantly improved the prediction accuracy resulting in lower prediction errors. The model seemed very promising for use in criminal investigation since it is based only on 16 loci (as a comparison most current DNA profiling methods are based on 15-16 STR loci as well); however, the proposed genome-wide techniques needed to obtain the methylation values require too much DNA for forensic use. Therefore, there was a strong need to develop methylation assays for the analysis of the selected 16 CpG sites that are both accurate and sensitive.

9.5 Next generation sequencing for age prediction

Next-generation sequencing technology has been adopted by the epigenetics community and its advantages compared to existing technologies have already been identified (Hurd & Nelson, 2009). Recently, a study tested the abilities of a bench-top sequencer (MiSeq®) in targeted methylation analysis of specific genomic regions (Masser *et al.*, 2013). Authors suggested that their proposed method was not only applicable to targeted DNA methylation studies but it could also be used to confirm

genome-wide studies, potentially overcoming technical issues faced in predominant approaches like Pyrosequencing®.

Adjusting a protocol that has specifically been developed for forensic SNP analysis, the methylation status of the proposed age-associated 16 CpG sites was successfully quantified. After validating the method in terms of linearity and reproducibility of methylation quantification and improving the methods based on experimental design, data from a total of 33 samples were used to test the age prediction accuracy using the ANN model [Figure 7-13]. Results were very encouraging in terms of sensitivity and accuracy of methylation quantification; however, age prediction accuracy was not as great (average absolute error of 13.32 years) as obtained when testing the ANN model using genome-wide methylation data. It is believed that there are mainly two possible explanations for this observation. Firstly, it should be noted that the data used to train the ANN model were obtained using a different methodology (Illumina's 27K and 450K Human methylation BeadChip technology (Bibikova *et al.*, 2011)). From this study, comparing the methylation detection between Pyrosequencing® and NGS, it is obvious that different methods could measure methylation in a different way; it may be consistent within the method but could vary between methods, since different types of 'bias' towards the detection of either the methylated or the unmethylated allele could be introduced. Therefore, it is believed that if the ANN model was trained on data obtained using the proposed next generation sequencing method, then the age prediction could potentially be better.

On the other hand, by testing two sets of DNA methylation standards amplified either using 30 or 45 PCR cycles, it was evident that amplification bias in bisulphite PCR could add variance [Figure 8-10]. This comes as no surprise as this type of bias had also been previously reported when testing tissue-specific differentially methylated CpG sites (Figure 5-5). 'Correcting' the obtained methylation levels was attempted by employing the equations of the observed vs. expected methylation levels of predefined DNA methylation controls. As a result, prediction accuracy was improved (average absolute error of 10.65 years), especially for older individuals. However, certainly for most markers, the PCR cycles could be further reduced since the resulting amplicon concentrations were much higher than the DNA amount needed for library preparation (only 1 µl out of 13 µl were used for each marker). Lastly, apart from bias

during the amplification of bisulphite-converted DNA, bias can be introduced post-PCR during the sequencing process as discussed in section 8.3.2.6. Adjusting the protocol to include fewer tube manipulations could possible reduce these bias.

Moreover, further optimisation of the protocol in terms of library input, library amplification and quantification could further improve the accuracy of methylation quantification and the distribution between markers and samples. Similar to the genome-wide DNA methylation data, MiSeq data can also undergo different normalisation strategies prior to analysis. Not only the 'linearity' graphs can be used for data correction for improved age prediction as shown in the present study, but also the data can be further corrected by using the average bisulfite conversion rates per amplicon for each sample. It is also believed that running samples in replicates (e.g. triplicates) would further reduce the prediction error by accounting for run-to-run variations in methylation detection, similar to the batch effects shown in the genome-wide platforms. In the future, running a larger number of samples would be vital before making any conclusions regarding the age prediction accuracy of the overall method.

Nevertheless, the results of this study are very encouraging as predicting age from an unknown sample looks possible (at least at the generation level at the moment). Although other forensic research groups have tried to predict age using other methods such as Pyrosequencing® (Zbieć-Piekarska *et al.*, 2015), it is believed that the MiSeq® will surpass the proposed current methods. As discussed in section 8.1.2, next generation sequencing (NGS) using MiSeq® is a very promising option for future forensics, since it can potentially incorporate STR typing (Warshauer *et al.*, 2013), mitochondrial genome sequencing (Parson *et al.*, 2013), Y-chromosome analysis (Xue *et al.*, 2009), differentiation of twins (Weber-Lehmann *et al.*, 2014), microbial forensics (Brenig *et al.*, 2010), species identification (Hajibabaei *et al.*, 2011), ancestry studies and phenotypic inferences (Yang *et al.*, 2014). Furthermore, the potential of NGS for tissue source identification was also proposed by a recent study (Bartling *et al.*, 2014).

9.6 Final remarks

As shown in this study, tissue- and age-specific DNA methylation profiling, together with tissue-specific mRNA profiling could be very useful in answer forensic questions, such as 'which body fluid/tissue is this crime scene stain/sample consists of?' or 'how old is the suspect?'. Although the results from the experiments included in this thesis are very promising and encouraging regarding these applications, it is believed that future experiments will further improve and strengthen the discussed conclusions. All the proposed tissue- and age-specific markers together with ones discovered in the future could be combined in multiplex PCR reactions (for example, a TissueID CpG multiplex and a MethAge multiplex) and be analysed in a simultaneous manner together with STR and SNP markers in an NGS platform. That way, information regarding the identity, ethnicity, age and physical characteristics of an individual together with the tissue source of the stain can be identified at the same time in a single experimental workflow. The forensic community is currently working through such a 'DNA witness' test and I believe that this thesis has successfully contributed towards this goal.

10 References

ABI (2014). Quantifiler Human and Y Human Male DNA Quantification kits - User Guide.

Akutsu, T., Motani, H., Watanabe, K., Iwase, H., Sakurada, K. (2012). Detection of bacterial 16S ribosomal RNA genes for forensic identification of vaginal fluid. *Legal Medicine (Tokyo)* **14(3),** 160-162.

Al-Qattan, S.I., Elfawal, M.A. (2010). Significance of teeth lead accumulation in age estimation. *Journal of Forensic and Legal Medicine* **17(6),** 325-358.

Alisch, R.S., Barwick, B.G., Chopra, P., Myrick, L.K., Satten, G.A., Conneely, K.N., Warren, S.T. (2012). Age-associated DNA methylation in pediatric populations. *Genome Research* **22(4),** 623-632.

Almen, M.S., Nilsson, E.K., Jacobsson, J.A., Kalnina, I., Klovins, J., Fredriksson, R., Schiöth, H.B. (2014). Genome-wide analysis reveals DNA methylation markers that vary with both age and obesity. *Gene* **548(1),** 61-67.

Alvarez, M., Juusola, J., Ballantyne, J. (2004). An mRNA and DNA co-isolation method for forensic casework samples. *Analytical Biochemistry* **335(2),** 289-298.

Ambion (2012). TURBO DNA-free kit.

Amiri, M., Derakhshandeh, K. (2011). *Applied artificial neural networks: from associative memories to biomedical applications*. InTech.

Amiri, M., Menhaj, M.B., Fallah, A. (2007). Stability analysis of self-feedback neural network structures. *Amirkabir Journal of Science and Technology* **14,** 103-109.

Ammerpohl, O., Martin-Subero, J.I., Richter, J., Vater, I., Siebert, R. (2009). Hunting for the 5th base: Techniques for analyzing DNA methylation. *Biochimica et Biophysica Acta* **1790(9),** 847-862.

An, J.H., Shin, K.J., Yang, W.I., Lee, H.Y. (2012). Body fluid identification in forensics. *BMB reports* **45(10),** 545-553.

An, J.H., Choi, A., Shin, K.J., Yang, W.I., Lee, H.Y. (2013). DNA methylation-specific multiplex assays for body fluid identification. *International Journal of Legal Medicine* **127(1),** 35-43.

Andréasson, H., Nilsson, M., Styrman, H., Pettersson, U., Allen, M. (2007). Forensic mitochondrial coding region analysis for increased discrimination using pyrosequencing technology. *Forensic Science International: Genetics* **1(1),** 35-43.

Anjum, S., Fourkala, E.O., Zikan, M., Wong, A., Gentry-Maharaj, A., Jones, A., Hardy, R., Cibula, D., Kuh, D., Jacobs, I.J., Teschendorff, A.E., Menon, U., Widschwendter, M. (2014). A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival. *Genome Medicine* **6(6),** 47.

Arany, S., Ohtani, S., Yoshioka, N., Gonmori, K. (2004). Age estimation from aspartic acid racemization of root dentin by internal standard method. *Forensic Science International* **141(2-3),** 127-130.

Bailey, A.J., Shimokomaki, M.S. (1971). Age-related changes in the reducible crosslinks of collagen. *FEBS Letters* **16(2),** 86-88.

Baker, D.J., Grimes, E.A., Hopwood, A.J. (2011). D-dimer assays for the identification of menstrual blood. *Forensic Science International* **212(1-3),** 210-214.

Baran, Y., Subramaniam, M., Biton, A., Tukiainen, T., Tsang, E.K., Rivas, M.A., Pirinen, M., Gutierrez-Arcelus, M., Smith, K.S., Kukurba, K.R., Zhang, R., Eng, C., Torgerson, D.G., Urbanek, C., GTEx Consortium, Li, J.B., Rodriguez-Santana, J.R., Burchard, E.G., Seibold, M.A., MacArthur, D.G., Montgomery, S.B., Zaitlen, N.A., Lappalainen, T. (2015). The landscape of genomic imprinting across diverse adult human tissues. *Genome Research* **25(7),** 927-936.

Barni, F., Lewis, S.W., Berti, A., Miskelly, G.M., Lago, G. (2007). Forensic application of the luminol reaction as a presumptive test for latent blood detection. *Talanta* **72(3),** 896-913.

Baron, U., Turbachova, I., Hellwag, A., Eckhardt, F., Berlin, K., Hoffmuller, U., Gardina, P., Olek, S. (2006). DNA methylation analysis as a tool for cell typing. *Epigenetics* **1(1),** 55-60.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116(2),** 281-297.

Bartling, C.M., Hester, M.E., Bartz, J., Heizer, E.Jr., Faith, S.A. (2014). Next-generation sequencing approach to epigenetic-based tissue source attribution. *Electrophoresis* **35(21-22),** 3096-3101.

Bartzeliotou, A.J., Dimitriadis, G.J. (1989). The state of DNA methylation of human e-globin gene in erythroid and non-erythroid cells. *Cell Differentiation and Development* **26(2),** 97-106.

Bauer, M., Patzelt, D. (2003). A method for simultaneous RNA and DNA isolation from dried blood and semen stains. *Forensic Science International* **136(1-3),** 76-78.

Bauer, M., Patzelt, D. (2008). Identification of menstrual blood by real time RT-PCR: technical improvements and the practical value of negative test results. *Forensic Science International* **174(1),** 55-59.

Behrend, L., Stöter, M., Kurth, M., Rutter, G., Heukeshoven, J., Deppert, W., Knippschild, U. (2000). Interaction of casein kinase 1 delta (CK1delta) with post-Golgi structures, microtubules and the spindle apparatus. *European Journal of Cell Biology* **79(4),** 240-251.

Bell, C.G., Teschendorff, A.E., Rakyan, V.K., Maxwell, A.P., Beck, S., Savage, D.A. (2010). Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. *BMC Medical Genomics* **3,** 33.

Bell, J.T., Loomis, A.K., Butcher, L.M., Gao, F., Zhang, B., Hyde, C.L., Sun, J., Wu, H., Ward, K., Harris, J., Scollen, S., Davies, M.N., Schalkwyk, L.C., Mill, J., MuTHER Consortium, Williams, F.M., Li, N., Deloukas, P., Beck, S., McMahon, S.B., Wang, J., John, S.L., Spector, T.D. (2014). Differential methylation of the TRPA1 promoter in pain sensitivity. *Nature Communications* **5**, 2978.

Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y. Pritchard, J.K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology* **12(1)**, R10.

Bell, J.T., Tsai, P.C., Yang, T.P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., Shin, S.Y., Dempster, E.L., Murray, R.M., Grundberg, E., Hedman, A.K., Nica, A., Small, K.S., MuTHER Consortium, Dermitzakis, E.T., McCarthy, M.I., Mill, J, Spector, T.D., Deloukas, P. (2012). Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genetics* **8(4)**, e1002629.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., Fan, J.B., Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics* **98(4)**, 288-295.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development* **16(1)**, 6-21.

Bird, A. (2007). Perceptions of epigenetics. *Nature* **447**, 396-398.

Bock, C. (2009). Epigenetic biomarker development. *Epigenomics* **1(1)**, 99-110.

Bocklandt, S., Lin, W., Sehl, M.E., Sanchez, F.J., Sinsheimer, J.S., Horvath, S., Vilain, E. (2011). Epigenetic predictor of age. *PLoS One* **6(6)**, e14821.

Boks, M.P., Derks, E.M., Weisenberger, D.J., Strengman, E., Janson, E., Sommer, I. E., Kahn, R.S., Ophoff, R.A. (2009). The relationship of DNA methylation with age, gender, genotype in twins and healthy controls. *PLoS One* **4(8)**, e6767.

Borghol, N., Blachére, T., Lefèvre, A. (2008). Transcriptional and epigenetic status of protamine 1 and 2 genes following round spermatids injection into mouse oocytes. *Genomics* **91(5)**, 415-422.

Bouakaze, C., Keyser, C., Crubezy, E., Montagnon, D., Ludes, B. (2009). Pigment phenotype and biogeographical ancestry from ancient skeletal remains: inferences from multiplexed autosomal SNP analysis. *International Journal of Legal Medicine* **123(4)**, 315-325.

Breit, T.M., Verschuren, M.C., Wolvers-Tettero, I.L., Van Gastel-Mol, E.J., Halhen, K., van Dongen, J.J. (1997). Human T cell leukemias with continuous V(D)J recombinase activity for TCR-delta gene deletion. *Journal of Immunology* **159(9)**, 4341-4349.

Brenig, B., Beck, J., Schütz, E. (2010). Shotgun metagenomics of biological stains using ultra-deep DNA sequencing. *Forensic Science International: Genetics* **4(4),** 228-231.

Britt-Compton, B., Rowson, J., Locke, M., Mackenzie, I., Kipling, D., Baird, D.M. (2006). Structural stability and chromosome-specific telomere length is governed by cis-acting determinants in humans. *Human Molecular Genetics* **15(5),** 725-733.

Bruder, C.E., Piotrowski, A., Gijsbers, A.A., Andersson, R., Erickson, S., Diaz de Ståhl, T., Menzel, U., Sandgren, J., von Tell, D., Poplawski, A., Crowley, M., Crasto, C., Partridge, E.C., Tiwari, H., Allison, D.B., Komorowski, J., van Ommen, G.J., Boomsma, D.I., Pedersen, N.L., den Dunnen, J.T., Wirdefeldt, K., Dumanski, J.P. (2008). Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *American Journal of Human Genetics* **82(3),** 763-771.

Butler, J.M. (2012). Advanced Topics in Forensic DNA Typing - Methodology. Elsevier.

Calonius, P.E. (1958). The leukocyte count in saliva. *Oral surgery, oral medicine, and oral pathology* **11(1),** 43-46.

Carlson, D.M. (1993). Salivary proline-rich proteins: biochemistry, molecular biology, and regulation of expression. *Critical reviews in oral biology and medicine* **4(3-4),** 495-502.

Carrette, O., Demalte, I., Scherl, A., Yalkinoglu, O., Corthals, G., Burkhard, P., Hochstrasser, D.F., Sanchez, J.C. (2003). A panel of cerebrospinal fluid potential biomarkers for the diagnosis of Alzheimer's disease. *Proteomics* **3(8),** 1486-1494.

Chaikind, B., Kilambi, K.P., Gray, J.J., Ostermeier, M. (2012). Targeted DNA methylation using an artificially bisected M.HhaI fused to zinc fingers. *PLoS One* **7(9),** e44852.

Chen, Y.A., Choufani, S., Ferreira, J.C., Grafodatskaya, D., Butcher, D.T., Weksberg, R. (2011). Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray. *Genomics* **97(4),** 214-222.

Choi, A., Shin, K.J., Yang, W.I., Lee, H.Y. (2014). Body fluid identification by integrated analysis of DNA methylation and body fluid-specific microbial DNA. *International Journal of Legal Medicine* **128(1),** 33-41.

Christensen, B.C., Houseman, E.A., Marsit, C.J., Zheng, S., Wrensch, M.R., Wiemels, J.L., Nelson, H.H., Karagas, M.R., Padbury, J.F., Bueno, R., Sugarbaker, D.J., Yeh, R.F., Wiencke, J.K., Kelsey, K.T. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genetics* **5(8),** e1000602.

Chu, P.G., Weiss, L.M. (2002). Keratin expression in human tissues and neoplasms. *Histopathology* **40(5),** 403-439.

Chu, Z.L., Wickrema, A., Krantz, J.C., Winkelmann, J.C. (1994). Erythroid-specific processing of human β spectrin I pre-mRNA. *Blood* **84(6),** 1992-1999.

Clark, S.J., Harrison, J., Paul, C.L., Frommer, M. (1994). High sensitivity mapping of methylated cytosines. *Nucleic Acids Research* **22(15),** 2990-2997.

Cosma, M.P., Tanaka, T., Nasmyth, K. (1999). Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. *Cell* **97(3),** 299-311.

Cunha, E., Baccino, E., Martrille, L., Ramsthaler, F., Prieto, J., Schuliar, Y., Lynnerup, N., Cattaneo, C. (2009). The problem of aging human remains and living individuals: a review. *Forensic Science International* **193(1-3),** 1-13.

Danielson, P.B. (2011). Isolation of highly specific protein markers for the identification of biological stains - Adapting comparative proteomics to forensics. U.S. Department of Justice, pp. 1-37.

Das, P.M., Singal, R. (2004). DNA methylation and cancer. *Journal of Clinical Oncology* **22(22),** 4632-4642.

Day, K., Waite, L.L., Thalacker-Mercer, A., West, A., Bamman, M.M., Brooks, J.D., Myers, R.M., Absher, D. (2013). Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biology* **14(9),** R102.

De Bustos, C., Ramos, E., Young, J.M., Tran, R.K., Menzel, U., Langford, C.F., Eichler, E.E., Hsu, L., Henikoff, S., Dumanski, J.P., Trask, B.J. (2009). Tissue-specific variation in DNA methylation levels along human chromosome 1. *Epigenetics & chromatin* **2(1),** 7.

de Lange, T. (2004). T-loops and the origin of telomeres. *Nature Reviews Molecular Cell Biology* **5(4),** 323-329.

Dehay, B., Martinez-Vicente, M., Ramirez, A., Perier, C., Klein, C., Vila, M., Bezard, E. (2012). Lysosomal dysfunction in Parkinson disease: ATP13A2 gets into the groove. *Autophagy* **8(9),** 1389-1391.

Dejeux, E., El abdalaoui, H., Gut, I.G., Tost, J. (2009). Identification and quantification of differentially methylated loci by the pyrosequencing technology. Second ed, pp. 189-205. Humana Press.

Divne, A.M., Edlund, H., Allen, M. (2010). Forensic analysis of autosomal STR markers using Pyrosequencing. *Forensic Science International: Genetics* **4(2),** 122-129.

Dobberstein, R.C., Huppertz, J., von Wurmb-Schwark, N., Ritz-Timme, S. (2008). Degradation of biomolecules in artificially and naturally aged teeth: Implications for age estimation on aspartic acid racemization and DNA analysis. *Forensic Science International* **179(2-3),** 181-191.

Dobberstein, R.C., Tung, S.M., Ritz-Timme, S. (2010). Aspartic acid racemisation in purified elastin from arteries as bases for age estimation. *International Journal of Legal Medicine* **124(4),** 269-275.

Dogde, J.E., Ramsahoye, B.H., Wo, Z.G., Okano, M., Li, E. (2002). De novo methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. *Gene* **289(1-2),** 41-48.

Dolce, V., Scarcia, P., Iacopetta, D., Palmieri, F. (2005). A fourth ADP/ATP carrier isoform in man: identification, bacterial expression, functional characterization and tissue distribution. *FEBS Letters* **579(3),** 633-637.

Donlin, L.T., Danzl, N.M., Wanjalla, C., Alexandropoulos, K. (2005). Deficiency in expression of the signaling protein Sin/Efs Leads to T lymphocyte activation and mucosal inflammation. *Molecular and Cellular Biology* **25(24),** 11035-11046.

Dupont, J.M., Tost, J., Jammes, H., Gut, I.G. (2004). De novo quantitative bisulfite sequencing using the pyrosequencing technology. *Analytical Biochemistry* **333(1),** 119-127.

Eads, C.A., Danenberg, K.D., Kawakami, K., Saltz, L.B., Blake, C., Shibata, D., Danenberg, P.V., Laird, P.W. (2000). MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Research* **28(8),** E32.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., Haefliger, C., Horton, R., Howe, K., Jackson, D.K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics* **38(12),** 1378-1385.

Edlund, H., Allen, M. (2009). Y chromosomal STR analysis using Pyrosequencing technology. *Forensic Science International: Genetics* **3(2),** 119-124.

El Hajj, N., Zechner, U., Schneider, E., Tresch, A., Gromoll, J., Hahn, T., Schorsch, M., Haaf, T. (2011). Methylation status of imprinted genes and repetitive elements in sperm DNA from infertile males. *Sexual Development* **5(2),** 60-69.

Espada, J., Esteller, M. (2010). DNA methylation and the functional organization of the nuclear compartment. *Seminars in Cell & Developmental Biology* **21(2),** 238-246.

Essex, M.J., Boyce, W.T., Hertzman, C., Lam, L.L., Armstrong, J.M., Neumann, S.M., Kobor, M.S. (2013). Epigenetic vestiges of early developmental adversity: childhood stress exposure and DNA methylation in adolescence. *Child Development* **84(1),** 58-75.

Eyre, D. (1987). Collagen cross-linking amino acids. *Methods in enzymology* **144,** 115-139.

Fang, R., Manohar, C.F., Shulse, C., Brevnov, M., Wong, A., Petrauskene, O.V., Brzoska, P., Furtado, M.R. (2006). Real-time PCR assays for the detection of tissue and body fluid specific mRNAs. *International Congress Series* **1288,** 685-687.

Fernandez-Jimenez, N., Plaza-Izurieta, L., Lopez-Euba, T., Jauregi-Miguel, A., Ramon Bilbao, J. (2012). Cubic regression-based degree of correction predicts the performance of whole bisulfitome amplified DNA methylation analysis. *Epigenetics* **7(12),** 1349-1354.

Fernandez, A.F., Huidobro, C., Fraga, M.F. (2012). De novo DNA methyltransferases: oncogenes, tumor suppressors, or both? *Trends in genetics* **28(10),** 474-479.

Fleming, R.I., Harbison, S. (2010a). The development of a mRNA multiplex RT-PCR assay for the definitive identification of body fluids. *Forensic Science International: Genetics* **4(4),** 244-256.

Fleming, R.I., Harbison, S. (2010b). The use of bacteria for the identification of vaginal secretions. *Forensic Science International: Genetics* **4(5),** 311-315.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suñer, D., Cigudosa, J.C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T.D., Wu, Y.Z., Plass, C., Esteller, M. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the U.S.A.* **102(30),** 10604-10609.

Fraga, M.F., Esteller, M. (2002). DNA methylation: A profile of methods and applications. *BioTechniques* **33,** 632-649.

Fragou, D., Fragou, A., Kouidou, S., Njau, S., Kovatsi, L. (2011). Epigenetic mechanisms in metal toxicity. *Toxicology Mechanisms and Methods* **21(4),** 343-352.

Fraser, I.S., McCarron, G., Markham, R., Resta, T. (1985). Blood and total fluid content of menstrual discharge. *Obstetrics and gynecology* **65(2),** 194-198.

Frazão, C., McVey, C.E., Amblar, M., Barbas, A., Vonrhein, C., Arraiano, C.M., Carrondo, M.A. (2006). Unravelling the dynamics of RNA degradation by ribonuclease II and its RNA-bound complex. *Nature* **443(7107),** 110-114.

Frumkin, D., Wasserstrom, A., Budowle, B., Davidson, A. (2011). DNA methylation-based forensic tissue identification. *Forensic Science International: Genetics* **5(5),** 517-524.

Frumkin, D., Wasserstrom, A., Davidson, A., Grafit, A. (2010). Authentication of forensic DNA samples. *Forensic Science International: Genetics* **4(2),** 95-103.

Fujimoto, S., Uratsuji, H., Saeki, H., Kagami, S., Tsunemi, Y., Komine, M., Tamaki, K. (2008). CCR4 and CCR10 are expressed on epidermal keratinocytes and are involved in cutaneous immune reaction. *Cytokine* **44(1),** 172-178.

Garagnani, P., Bacalini, M.G., Pirazzini, C., Gori, D., Giuliani, C., Mari, D., Di Blasio, A.M., Gentilini, D., Vitale, G., Collino, S., Rezzi, S., Castellani, G., Capri, M., Salvioli, S., Franceschi, C. (2012). Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell* **11(6),** 1132-1134.

Gentilini, D., Mari, D., Castaldi, D., Remondini, D., Ogliari, G., Ostan, R., Bucci, L., Sirchia, S.M., Tabano, S., Cavagnini, F., Monti, D., Franceschi, C., Di Blasio, A.M., Vitale, G. (2013). Role of epigenetics in human aging and longevity: genome-wide DNA methylation profile in centenarians and centenarians' offspring. *Age (Dordrecht, Netherlands)* **35(5),** 1961-1973.

Gentry, L.E., Thacker, M.A., Doughty, R., Timkovich, R., Busenlehner, L.S. (2013). His86 from the N-terminus of frataxin coordinates iron and is required for Fe-S cluster synthesis. *Biochemistry* **52(35),** 6085-6096.

Giampaoli, S., Berti, A., Valeriani, F., Gianfranceschi, G., Piccolella, A., Buggiotti, L., Rapone, C., Valentini, A., Ripani, L., Romano Spica, V. (2012). Molecular identification of vaginal fluid by microbial signature. *Forensic Science International: Genetics* **6(5),** 559-564.

Gibbs, S., Fijneman, R., Wiegant, J., van Kessel, A.G., van De Putte, P., Backendorf, C. (1993). Molecular characterisation and evolution of the SPRR family of keratinocyte differentiation markers encoding small proline-rich proteins. *Genomics* **16(3),** 630-637.

Gilder, J., Koppl, R., Kornfield, I., Krane, D., Mueller, L., Thompson, W. (2009). Comments on the review of low copy number testing. *International Journal of Legal Medicine* **123,** 535-536.

Gipson, I.K., Ho, S.B., Spurr-Michaud, S.J., Tisdale, A.S., Zhan, Q., Torlakovic, E., Pudney, J., Anderson, D.J., Toribara, N.W., Hill, J.A. 3rd. (1997). Mucin genes expressed by human female reproductive tract epithelia. *Biology of reproduction* **56(4),** 999-1011.

Glass, D., Vinuela, A., Davies, M.N., Ramasamy, A., Parts, L., Knowles, D., Brown, A.A., Hedman, A.K., Small, K.S., Buil, A., Grundberg, E., Nica, A.C., Di Meglio, P., Nestle, F.O., Ryten, M., UK Brain Expression consortium, MuTHER consortium, Durbin, R., McCarthy, M.I., Deloukas, P., Dermitzakis, E.T., Weale, M.E., Bataille, V., Spector, T.D. (2013). Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biology* **14(7),** R75.

Goffin, F., Munaut, C., Frankenne, F., Perrier D'Hauterive, S., Beliard, A., Fridman, V., Nervo, P., Colige, A., Foidart, J.M. (2003). Expression pattern of metalloproteinases and tissue inhibitors of matrix-metalloproteinases in cycling human endometrium. *Biology of reproduction* **69(3),** 976-984.

Gomes, I., Kohlmeier, F., Schneider, P.M. (2011). Genetic markers for body fluid and tissue identification in forensics. *Forensic Science International: Genetics Supplement Series* **3(1),** e469-e470.

Gomes, I., Strohbücker, B., Rothschild, M.A., Schneider, P.M. (2013). Evaluation of mRNA specific markers using a pentaplex system for the identification of skin and saliva from contact trace evidence. *Forensic Science International: Genetics Supplement Series* **4(1),** e180-e181.

Gonzalgo, M.L., Jones, P.A. (2002). Quantitative methylation analysis using methylation-sensitive single-nucleotide primer extension (Ms-SNuPE). *Methods* **27(2),** 128-133.

Gray, D., Frascione, N., Daniel, B. (2012). Development of an immunoassay for the differentiation of menstrual blood from peripheral blood. *Forensic Science International* **220(1-3),** 12-18.

Grönniger, E., Weber, B., Heil, O., Peters, N., Stäb, F., Wenck, H., Korn, B., Winnefeld, M., Lyko, F. (2010). Aging and chronic sun exposure cause distinct epigenetic changes in human skin. *PLoS Genetics* **6(5),** e1000971.

Gubin, A.N., Miller, J.L. (2001). Human erythroid porphobilinogen deaminase exists in 2 splice variants. *Blood* **97(3),** 815-817.

Guo, J.H., Maltha, J.C., He, S.G., Krapels, I.P., Spauwen, P.H., Steegers-Theunissen, R.P., Von den Hoff, J.W. (2006). Cytokeratin expression in palatal and marginal mucosa of cleft palate patients. *Archives of oral biology* **51(7),** 573-580.

Haas, C., Hanson, E., Anjos, M.J., Ballantyne, K.N., Banemann, R., Bhoelai, B., Borges, E., Carvalho, M., Courts, C., De Cock, G., Drobnic, K., Dötsch, M., Fleming, R., Franchi, C., Gomes, I., Hadzic, G., Harbison, S.A., Harteveld, J., Hjort, B., Hollard, C., Hoff-Olsen, P., Hüls, C., Keyser, C., Maroñas, O., McCallum, N., Moore, D., Morling, N., Niederstätter, H., Noël, F., Parson, W., Phillips, C., Popielarz, C., Roeder, A.D., Salvaderi, L., Sauer, E., Schneider, P.M., Shanthan, G., Syndercombe Court, D., Turanská, M., van Oorschot, R.A., Vennemann, M., Vidaki, A., Zatkalíková, L., Ballantyne, J. (2014). RNA/DNA co-analysis from human menstrual blood and vaginal secretion stains: Results of a fourth and fifth collaborative EDNAP exercise. *Forensic Science International: Genetics* **8(1),** 203-212.

Haas, C., Hanson, E., Anjos, M.J., Banemann, R., Berti, A., Borges, E., Carracedo, A., Carvalho, M., Courts, C., De Cock, G., Dötsch, M., Flynn, S., Gomes, I., Hollard, C., Hjort, B., Hoff-Olsen, P., Hríbiková, K., Lindenbergh, A., Ludes, B., Maroñas, O., McCallum, N., Moore, D., Morling, N., Niederstätter, H., Noel, F., Parson, W., Popielarz, C., Rapone, C., Roeder, A.D., Ruiz, Y., Sauer, E., Schneider, P.M., Sijen, T., Syndercombe Court, D., Sviežená, B., Turanská, M., Vidaki, A., Zatkalíková, L., Ballantyne, J. (2013a). RNA/DNA co-analysis from human saliva and semen stains--results of a third collaborative EDNAP exercise. *Forensic Science International: Genetics* **7(2),** 230-239.

Haas, C., Hanson, E., Anjos, M.J., Bar, W., Banemann, R., Berti, A., Borges, E., Bouakaze, C., Carracedo, A., Carvalho, M., Castella, V., Choma, A., De Cock, G., Dötsch, M., Hoff-Olsen, P., Johansen, P., Kohlmeier, F., Lindenbergh, P.A., Ludes, B., Maroñas, O., Moore, D., Morerod, M.L., Morling, N., Niederstätter, H., Noel, F., Parson, W., Patel, G., Popielarz, C., Salata, E., Schneider, P.M., Sijen, T., Sviežena, B., Turanská, M., Zatkalíková, L., Ballantyne, J. (2012). RNA/DNA co-analysis from blood stains: results of a second collaborative EDNAP exercise. *Forensic Science International: Genetics* **6(1),** 70-80.

Haas, C., Hanson, E., Ballantyne, J. (2013b). mRNA and MicroRNA for Body Fluid Identification, *Encyclopedia of Forensic Sciences,* 2nd edition, pp. 402-408.

Haas, C., Hanson, E., Banemann, R., Bento, A.M., Berti, A., Carracedo, A., Courts, C., De Cock, G., Drobnic, K., Fleming, R., Franchi, C., Gomes, I., Hadzic, G., Harbison, S.A., Hjort, B., Hollard, C., Hoff-Olsen, P., Keyser, C., Kondili, A., Maroñas, O., McCallum, N., Miniati, P., Morling, N., Niederstätter, H., Noël, F., Parson, W., Porto, M.J., Roeder, A.D., Sauer, E., Schneider, P.M., Shanthan, G., Sijen, T., Syndercombe Court, D., Turanská, M., van den Berge, M., Vennemann, M., Vidaki, A., Zatkalíková, L., Ballantyne, J. (2015). RNA/DNA co-analysis from human skin and contact traces - results of a sixth collaborative EDNAP exercise. *Forensic Science International: Genetics* **16,** 139-147.

Haas, C., Hanson, E., Bär, W., Banemann, R., Bento, A. M., Berti, A., Borges, E., Bouakaze, C., Carracedo, A., Carvalho, M., Choma, A., Dötsch, M., Duriancikova, M., Hoff-Olsen, P., Hohoff, C., Johansen, P., Lindenbergh, P.A., Loddenkötter, B., Ludes, B., Maroñas, O., Morling, N., Niederstätter, H., Parson, W., Patel, G., Popielarz, C., Salata, E., Schneider, P.M., Sijen, T., Sviezená, B., Zatkalíková, L., Ballantyne, J. (2011a). mRNA profiling for the identification of blood--results of a collaborative EDNAP exercise. *Forensic Science International: Genetics* **5(1),** 21-26.

Haas, C., Hanson, E., Kratzer, A., Bär, W., Ballantyne, J. (2011b). Selection of highly specific and sensitive mRNA biomarkers for the identification of blood. *Forensic Science International: Genetics* **5(5),** 449-458.

Haas, C., Klesser, B., Maake, C., Bär, W., Kratzer, A. (2009a). mRNA profiling for body fluid identification by reverse transcription endpoint PCR and realtime PCR. *Forensic Science International: Genetics* **3(2),** 80-88.

Haas, C., Muheim, C., Kratzer, A., Bär, W., Maake, C. (2009b). mRNA profiling for the identification of sperm and seminal plasma. *Forensic Science International: Genetics Supplement Series* **2(1),** 534-535.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A., Baird, D.J. (2011). Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One* **6(4),** e17497.

Hammoud, S.S., Purwar, J., Pflueger, C., Cairns, B.R., Carrell, D.T. (2010). Alterations in sperm DNA methylation patterns at imprinted loci in two classes of infertility. *Fertility and sterility* **94(5),** 1728-1733.

Han, X., Chesney, R.W. (2003). Regulation of taurine transporter gene (TauT) by WT1. *FEBS Letters* **540(1-3),** 71-76.

Han, X., Patters, A.B., Jones, D.P., Zelikovic, I., Chesney, R.W. (2006). The taurine transporter: mechanisms of regulation. *Acta physiologica* (Oxford, England) **187(1-2),** 61-73.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell* **49(2),** 359-367.

Hanson, E., Haas, C., Jucker, R., Ballantyne, J. (2011). Identification of skin in touch/contact forensic samples by messenger RNA profiling. *Forensic Science International: Genetics Supplement Series* **3(1),** e305-e306.

Hanson, E., Haas, C., Jucker, R., Ballantyne, J. (2012). Specific and sensitive mRNA biomarkers for the identification of skin in 'touch DNA' evidence. *Forensic Science International: Genetics* **6(5),** 548-558.

Hanson, E.K., Ballantyne, J. (2013a). Highly specific mRNA biomarkers for the identification of vaginal secretions in sexual assault investigations. *Science & Justice* **53(1),** 14-22.

Hanson, E.K., Ballantyne, J. (2013b). Multiplex high resolution melt (HRM) messenger RNA profiling assays for body fluid identification. *Forensic Science International: Genetics Supplement Series* **4(1),** e125-e126.

Harteveld, J., Lindenbergh, A., Sijen, T. (2013). RNA cell typing and DNA profiling of mixed samples: can cell types and donors be associated? *Science & Justice* **53(3),** 261-269.

Herman, J.G., Graff, J.R., Myohanen, S., Nelkin, B.D., Baylin, S.B. (1996). Methylation-specific PCR: A novel PCR assay for methylation status of CpG islands. *Proceedings of the National Academy of Sciences of the U.S.A.* **93(18),** 9821-9826.

Heyn, H., Li, N., Ferreira, H.J., Moran, S., Pisano, D.G., Gomez, A., Diez, J., Sanchez-Mut, J.V., Setien, F., Carmona, F.J., Puca, A.A., Sayols, S., Pujana, M.A., Serra-Musach, J., Iglesias-Platas, I., Formiga, F., Fernandez, A.F., Fraga, M.F., Heath, S.C., Valencia, A., Gut, I.G., Wang, J., Esteller, M. (2012). Distinct DNA methylomes of newborns and centenarians. *Proceedings of the National Academy of Sciences of the U.S.A.* **109(26),** 10522-10527.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology* **14(10),** R115.

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R. S., Boks, M. P., van Eijk, K., van den Berg, L.H., Ophoff, R.A. (2012). Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biology* **13(10),** R97.

Hsu, C.C., Leu, Y.W., Tseng, M.J., Lee, K.D., Kuo, T.Y., Yen, J.Y., Lai, Y.L., Hung, Y.C., Sun, W.S., Chen, C.M., Chu, P.Y., Yeh, K.T., Yan, P.S., Chang, Y.S., Huang, T.H., Hsiao, S.H. (2011). Functional characterization of Trip10 in cancer cell growth and survival. *Journal of Biomedical Science* **18,** 12.

Human Tissue Act (HTA) (2014). Code of practice 1 - Consent (H. T. Authority, Ed.).

Hua, D., Hu, Y., Wu, Y.Y., Cheng, Z.H., Yu, J., Du, X., Huang, Z.H. (2011). Quantitative methylation analysis of multiple genes using methylation-sensitive restriction enzyme-based quantitative PCR for the detection of hepatocellular carcinoma. *Experimental and molecular pathology* **91(1),** 455-460.

Huang, D., Lin, X., Chen, H., Yang, Q., Jie, Y., Zhai, X., Yin, H. (2008). Parentally imprinted allele (PIA) typing in the differentially methylated region upstream of the human H19 gene. *Forensic Science International: Genetics* **2(4),** 286-291.

Huang, J.T., Leweke, F.M., Oxley, D., Wang, L., Harris, N., Koethe, D., Gerth, C.W., Nolden, B.M., Gross, S., Schreiber, D., Reed, B., Bahn, S. (2006). Disease biomarkers in cerebrospinal fluid of patients with first-onset psychosis. *PLoS Medicine* **3(11),** e428.

Hurd, P.J., Nelson, C.J. (2009). Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings in Functional Genomics & Proteomics* **8(3),** 174-183.

Illingworth, R., Kerr, A., DeSousa, D., Jorgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., Humphray, S., Cox, T., Langford, C., Bird, A. (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PloS Biology* **6(1),** e22.

Illumina (2013a). An Introduction to Next-Generation Sequencing Technology.

Illumina (2013b). MiSeq System User Guide.

Invitrogen (2010). Qubit dsDNA HS Assay kits.

Invitrogen (2013). Superscript III FirstStrand Synthesis system for RT-PCR.

Ishikawa, N., Hara, J., Takasaka, T., Kobayashi, M., Yoshimoto, K., Ikegaya, H. (2011). Age estimation using cytochrome c oxidase activity analysis. *Forensic Science International* **209(1-3),** 48-52.

Ishino, M., Ohba, T., Inazawa, J., Sasaki, H., Ariyama, Y., Sasaki, T. (1997). Identification of an Efs isoform that lacks the SH3 domain and chromosomal mapping of human Efs. *Oncogene* **15(14),** 1741-1745.

Jacinto, F.V., Ballestar, E., Esteller, M. (2008). Methyl-DNA immunoprecipitation (MeDIP): Hunting down the DNA methylome. *BioTechniques* **44(1),** 35-39.

Jackson, B., Tilli, C.M., Hardman, M.J., Avillion, A.A., MacLeod, M.C., Ashcroft, G.S., Byrne, C. (2005). Late cornified envelope family in differentiating epithelia-response to calcium and untraviolet irradiation. *The Journal of Investigative Dermatology* **124(5),** 1062-1070.

Jakubowska, J., Maciejewska, A., Pawlowski, R., Bielawski, K.P. (2013). mRNA profiling for vaginal fluid and menstrual blood identification. *Forensic Science International: Genetics* **7(2),** 272-278.

Jiang, H., Ju, Z., Rudoplh, K.L. (2007). Telomere shortening and ageing. *Journal of Gerontology and Geriatry* **40(5),** 314-324.

Johansson, A., Enroth, S., Gyllensten, U. (2013). Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. *PLoS One* **8(6),** e67378.

Jones Jr., E.L. (2005). The identification of semen and other body fluids. Forensic Science Handbook, pp. 329-382. Prentice Hall, Upper Saddle River, NJ.

Jones, M.J., Farre, P., McEwen, L.M., Maclsaac, J.L., Watt, K., Neumann, S.M., Emberly, E., Cynader, M.S., Virji-Babul, N., Kobor, M.S. (2013). Distinct DNA methylation patterns of cognitive impairment and trisomy 21 in down syndrome. *BMC Medical Genomics* **6,** 58.

Jumpertz, R., Hanson, R.L., Sievers, M.L., Bennett, P.H., Nelson, R.G., Krakoff, J. (2011). Higher energy expenditure in humans predicts natural mortality. *The Journal of Clinical Endocrinology and Metabolism* **96(6),** E972-E976.

Juusola, J., Ballantyne, J. (2003). Messenger RNA profiling: a prototype method to supplant conventional methods for body fluid identification. *Forensic Science International* **135(2),** 85-96.

Juusola, J., Ballantyne, J. (2005). Multiplex mRNA profiling for the identification of body fluids. *Forensic Science International* **152(1),** 1-12.

Kalinin, A., Marekov, L.N., Steinert, P.M. (2001). Assembly of the epidermal cornified cell envelope. *Journal of Cell Science* **114(17),** 3069-3070.

KAPA Biosystems (2014). KAPA Library Quantification Kits for Illumina sequencing platforms.

Kendall Harden, T., Boyer, J.L., Nicholas, R.A. (1995). P2-Purinergic receptors: subtype-associated signaling responses and structure *Annual Review of Pharmacoloogy and Toxicology* **35,** 541-579.

Kobus, H.J., Silenieks, E., Scharnberg, J. (2002). Improving the effectiveness of fluorescence for the detection of semen stains on fabrics. *Journal of Forensic Sciences* **47(4),** 819-823.

Koch, C.M., Joussen, S., Schellenberg, A., Lin, Q., Zenke, M., Wagner, W. (2012). Monitoring of cellular senescence by DNA-methylation at specific CpG sites. *Aging Cell* **11(2),** 366-369.

Koch, C.M., Suschek, C.V., Lin, Q., Bork, S., Goergens, M., Joussen, S., Pallua, N., Ho, A.D., Zenke, M., Wagner, W. (2011). Specific age-associated DNA methylation changes in human dermal fibroblasts. *PLoS One* **6(2),** e16679.

Koch, C.M., Wagner, W. (2011). Epigenetic-aging-signature to determine age in different tissues. *Aging (Albany NY)* **3(10),** 1018-1027.

Kohlmeier, F., Schneider, P.M. (2012). Successful mRNA profiling of 23 years old blood stains. *Forensic Science International: Genetics* **6(2),** 274-276.

Kovatsi, L., Georgiou, E., Ioannou, A., Haitoglou, C., Tzimagiorgis, G., Tsoukali, H., Kouidou, S. (2010). p16 promoter methylation in Pb^{2+} -exposed individuals. *Clinical Toxicology (Philadelphia)* **48(2),** 124-128.

Laaksonen, M., Kaliste-Korhonen, E., Karenlampi, S., Hanninen, O. (1995). P450 enzyme CYP2B catalyses the detoxification of diisopropyl fluorophosphate. *Chemico-Biological Interactions* **94(3),** 197-213.

Lackner, J.E., Herwig, R., Schmidbauer, J., Schatzl, G., Kratzik, C., Marberger, M. (2006). Correlation of leukocytospermia with clinical infection and the positive effect of antiinflammatory treatment on semen quality. *Fertility and Sterility* **86(3),** 601-605.

LaRue, B.L., King, J.L., Budowle, B. (2013). A validation study of the Nucleix DSI-Semen kit--a methylation-based assay for semen identification. *International Journal of Legal Medicine* **127(2),** 299-308.

Lasken, R.S., Egholm, M. (2003). Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. *Trends in Biotechnology* **21(12),** 531-535.

Lawton, K.A., Berger, A., Mitchell, M., Milgram, K.E., Evans, A.M., Guo, L., Hanson, R.W., Kalhan, S.C., Ryals, J.A., Milburn, M.V. (2008). Analysis of the adult human plasma metabolome. *Pharmacogenomics* **9(4),** 383-397.

Lee, H.Y., Park, M.J., Choi, A., An, J.H., Yang, W.I., Shin, K.J. (2012a). Potential forensic application of DNA methylation profiling to body fluid identification. *International Journal of Legal Medicine* **126(1),** 55-62.

Lee, K.W., Pausova, Z. (2013). Cigarette smoking and DNA methylation. *Frontiers in Genetics* **4,** 132.

Lee, P.Y., Costumbrado, J., Hsu, C.Y., Kim, Y.H. (2012b). Agarose gel electrophoresis for the separation of DNA fragments. *Journal of* Visualised *Experiments* **Apr 20**, 62.

Levings, P.P., Bungert, J. (2002). The human β-globin locus control region. *European Journal of Biochemistry* **269(6),** 1589-1599.

Li, C., Zhang, S., Que, T., Li, L., Zhao, S. (2011). Identical but not the same: The value of DNA methylation profiling in forensic discrimination within monozygotic twins. *Forensic Science International: Genetics Supplement Series* **3(1),** e337-e338.

Li, C., Zhao, S., Zhang, N., Zhang, S., Hou, Y. (2013). Differences of DNA methylation profiles between monozygotic twins' blood samples. *Molecular Biology Reports* **40(9),** 5275-5280.

Li, E., Beard, C., Jaenisch, R. (1993). Role of DNA methylation in genomic imprinting. *Nature* **366(6453),** 362-365.

Li, H., Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics* **26(5),** 589-595.

Li, H., Handsaker, B., Wysoker, A., Fennel, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25(16),** 2078-2079.

Li, S.X., Liu, L.J., Jiang, W.G., Sun, L.L., Zhou, S.J., Le Foll, B., Zhang, X.Y., Kosten, T.R., Lu, L. (2010). Circadian alteration in neurobiology during protracted opiate withdrawal in rats. *Journal of Neurochemistry* **115(2),** 353-362.

Liang, Q., Liu, L., Peng, J., Sun, Z., Jiang, S. (2008). Determination and Difference Analysis of DNA Methylation Content Both in Blood and Muscle Tissue of Pigs. *Agricultural Sciences in China* **7(8),** 1010-1015.

Life Technologies (2014). DNA Fragment Analysis by Capillary Electrophoresis.

Lindenbergh, A., de Pagter, M., Ramdayal, G., Visser, M., Zubakov, D., Kayser, M., Sijen, T. (2012). A multiplex (m)RNA-profiling system for the forensic identification of body fluids and contact traces. *Forensic Science International: Genetics* **6(5),** 565-577.

Lindenbergh, A., Maaskant, P., Sijen, T. (2013). Implementation of RNA profiling in forensic casework. *Forensic Science International: Genetics* **7(1),** 159-166.

Liu, B., Lague, J.R., Nunes, D.P., Toseli, P., Oppenheim, F.G., Soares, R.V., Troxler, R.F., Offner, G.D. (2002). Expression of membrane-associated mucins MUC1 and MUC4 in major human salivary glands. *The Journal of Histochemistry and Cytochemistry* **50(6),** 811-820.

Liu, J., Morgan, M., Hutchison, K., Calhoun, V.D. (2010). A study of the influence of sex on genome wide methylation. *PLoS One* **5(4),** e10028.

Lowe, A., Murray, C., Whitaker, J., Tully, G., Gill, P. (2002). The propensity of individuals to deposit DNA and secondary transfer of low level DNA from individuals to inert surfaces. *Forensic Science International* **129(1),** 25-34.

Lowenson, J., Clarke, S. (1988). Does the chemical instability of aspartyl and asparaginyl residues in proteins contribute to erythrocyte aging? The role of protein carboxyl methylation reactions. *Blood Cells* **14(1),** 103-118.

Lu, S., Davies, P.J. (1997). Regulation of the expression of the tissue transglutaminase gene by DNA methylation. *Proceedings of the National Academy of Sciences of the U.S.A.* **94(9),** 4692-4697.

Lu, T., Pan, Y., Kao, S.Y., Li, C., Kohane, I., Chan, J., Yankner, B.A. (2004). Gene regulation and DNA damage in the ageing human brain. *Nature* **429,** 883-891.

Lynnerup, N., Kjeldsen, H., Zweihoff, R., Heegaard, S., Jacobsen, C., Heinemeier, J. (2010). Ascertaining year of birth/age at death in forensic cases: A review of conventional methods and methods allowing for absolute chronology. *Forensic Science International* **201(1-3),** 74-78.

Ma, L.L., Yi, S.H., Huang, D.X., Mei, K., Yang, R.Z. (2013). Screening and identification of tissue-specific methylation for body fluid identification. *Forensic Science International: Genetics Supplement Series* **4(1),** e37-e38.

Macfarlan, T., Parker, J.B., Nagata, K., Chakravarti, D. (2006). Thanatos-associated protein 7 associates with template activating factor-Ibeta and inhibits histone acetylation to repress transcription. *Molecular Endocrinology* **20(2),** 335-347.

Madi, T., Balamurugan, K., Bombardi, R., Duncan, G., McCord, B. (2012). The determination of tissue-specific DNA methylation patterns in forensic biofluids using bisulfite modification and pyrosequencing. *Electrophoresis* **33(12),** 1736-1745.

Maeda, H., Ishikawa, T., Michiue, T. (2011). Forensic biochemistry for functional investigation of death: concept and practical application. *Legal Medicine (Tokyo)* **13(2),** 55-67.

Maeda, H., Zhu, B.L., Ishikawa, T., Michiue, T. (2010). Forensic molecular pathology of violent deaths. *Forensic Science International* **203(1-3),** 83-92.

Maia, M., de Vriese, A., Janssens, T., Moons, M., van Landuyt, K., Tavernier, J., Lories, R.J., Conway, E.M. (2010). CD248 and its cytoplasmic domain: a therapeutic target for arthritis. *Arthritis and Rheumatism* **62(12),** 3595-3606.

Manabe, F., Tsutsumi, A., Yamamoto, Y., Hashimoto, Y., Ishizu, H. (1991). Identification of human semen by a chemiluminescent assay of choline. *Japanese Journal of Legal Medicine* **45(3),** 205-215.

Mandavilli, B.S., Santos, J.H., van Houten, B. (2002). Mitochondrial DNA repair and aging. *Mutation Research* **509(1-2),** 127-151.

Martin-de las Heras, S., Valenzuela, A., Villanueva, E. (1999). Deoxypyridinoline crosslinks in human dentin and estimation of age. *International Journal of Legal Medicine* **112(4),** 222-226.

Martin-Subero, J.I., Kreuz, M., Bibikova, M., Bentink, S., Ammerpohl, O., Wickham-Garcia, E., Rosolowski, M., Richter, J., Lopez-Serra, L., Ballestar, E., Berger, H., Agirre, X., Bernd, H.W., Calvanese, V., Cogliatti, S.B., Drexler, H.G., Fan, J.B., Fraga, M.F., Hansmann, M.L., Hummel, M., Klapper, W., Korn, B., Küppers, R., Macleod, R.A., Möller, P., Ott, G., Pott, C., Prosper, F., Rosenwald, A., Schwaenen, C., Schübeler, D., Seifert, M., Stürzenhofecker, B., Weber, M., Wessendorf, S., Loeffler, M., Trümper, L., Stein, H., Spang, R., Esteller, M., Barker, D., Hasenclever, D., Siebert, R., Molecular Mechanisms in Malignant Lymphomas Network Project of the Deutsche Krebshilfe. (2009). New insights into the biology and origin of mature aggressive B-cell lymphomas by combined epigenomic, genomic, and transcriptional profiling. *Blood* **113(11),** 2488-2497.

Martin, N.C., Clayson, N.J., Scrimger, D.G. (2006). The sensitivity and specificity of red-starch paper for the detection of saliva. *Science & Justice* **46(2),** 97-105.

Martino, D., Loke, Y.J., Gordon, L., Ollikainen, M., Cruickshank, M.N., Saffery, R., Craig, J.M. (2013). Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance. *Genome Biology* **14(5),** R42.

Masser, D.R., Berg, A.S., Freeman, W.M. (2013). Focused, high accuracy 5-methylcytosine quantitation with base resolution by benchtop next-generation sequencing. *Epigenetics & Chromatin* **6(1),** 33.

Matheson, C.D., Veall, M.A. (2014). Presumptive blood test using Hemastix® with EDTA in archaeology. *Journal of Archaeological Science* **41,** 230-241.

McKenna, A., Hanna, M., Banks, E., Sivanchenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altschuler, D., Gabriel, S., Daly, M., DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20(9),** 1297-1303.

Meissner, C., Bruse, P., Oehmichen, M. (2006). Tissue-specific deletion patterns of the mitochondrial genome with advancing age. *Experimental Gerontology* **41(5),** 518-524.

Meissner, C., Ritz-Timme, S. (2010). Molecular pathology and age estimation. *Forensic Science International* **203(1-3),** 34-43.

Meissner, C., von Wurmb, N., Oehmichen, M. (1997). Detection of the age-dependent 4977 bp deletion of mitochondrial DNA. A pilot study. *International Journal of Legal Medicine* **110(5),** 288-291.

Meissner, C., von Wurmb, N., Schimansky, B., Oehmichen, M. (1999). Estimation of age at death based on quantitation of the 4977-bp deletion of human mitochondrial DNA in skeletal muscle. *Forensic Science International* **105(2),** 115-124.

Menni, C., Kastenmuller, G., Petersen, A.K., Bell, J.T., Psatha, M., Tsai, P.C., Gieger, C., Schulz, H., Erte, I., John, S., Brosnan, M.J., Wilson, S.G., Tsaprouni, L., Lim, E.M., Stuckey, B., Deloukas, P., Mohney, R., Suhre, K., Spector, T.D., Valdes, A.M. (2013). Metabolomic markers reveal novel pathways of ageing and early development in human populations. *International Journal of Epidemiology* **42(4),** 1111-1119.

Mohamed, S.A., Wesch, D., Blumenthal, A., Bruse, P., Windler, K., Ernst, M., Kabelitz, D., Oehmichen, M., Meissner, C. (2004). Detection of the 4977 bp deletion of mitochondrial DNA in different human blood cells. *Experimental Gerontology* **39(2),** 181-188.

Moreno, L.I., Tate, C.M., Knott, E.L., McDaniel, J.E., Rogers, S.S., Koons, B.W., Kavlick, M.F., Craig, R.L., Robertson, J.M. (2012). Determination of an effective housekeeping gene for the quantification of mRNA for forensic applications. *Journal of Forensic Sciences* **57(4),** 1051-1058.

Moskalev, E.A., Zavgorodnij, M.G., Majorova, S.P., Vorobjev, I.A., Jandaghi, P., Bure, I.V., Hoheisel, J.D. (2011). Correction of PCR-bias in quantitative DNA methylation studies by means of cubic polynomial regression. *Nucleic Acids Research* **39(11),** e77.

Naito, E., Dewa, K., Fukuda, M., Sumi, H., Wakabayashi, Y., Umetsu, K., Yuasa, I., Yamanouchi, H. (2003). Novel paternity testing by distinguishing parental alleles at a VNTR locus in the differentially methylated region upstream of the human H19 gene. *Journal of Forensic Sciences* **48(6),** 1275-1279.

Naito, E., Dewa, K., Yamanouchi, H., Takagi, S., Kominami, R. (1993). Sex determination using the hypomethylation of a human macro-satellite DXZ4 in female cells. *Nucleic Acids Research* **21(10),** 2533-2534.

Nakatome, M., Orii, M., Hamajima, M., Hirata, Y., Uemura, M., Hirayama, S., Isobe, I. (2011). Methylation analysis of circadian clock gene promoters in forensic autopsy specimens. *Legal Medicine (Tokyo)* **13(4),** 205-209.

Nakayashiki, N., Takamiya, M., Shimamoto, K., Aoki, Y. (2009a). Analysis of the methylation profiles in imprinted genes applicable to parental allele discrimination. *Legal Medicine (Tokyo)* **11 Suppl 1,** S471-S472.

Nakayashiki, N., Takamiya, M., Shimamoto, K., Aoki, Y., Hashiyada, M. (2008). Studies on differentially methylated parental allele in imprinted genes. *Forensic Science International: Genetics Supplement Series* **1(1),** 572-573.

Nakayashiki, N., Takamiya, M., Shimamoto, K., Aoki, Y., Hashiyada, M. (2009b). Investigation of the methylation status around parent-of-origin detectable SNPs in imprinted genes. *Forensic Science International: Genetics* **3(4),** 227-232.

Nam, H., Whang, K., Lee, Y. (2007). Analysis of vaginal lactic acid producing bacteria in healthy women. *Journal of Microbiology* **45(6),** 515-520.

National DNA Database Strategy Board (2012-2013). Annual Report (Home Office, Ed.), pp. 1-26.

Neumann, L.C., Weinhausel, A., Thomas, S., Horsthemke, B., Lohmann, D.R., Zeschnigk, M. (2011). EFS shows biallelic methylation in uveal melanoma with poor prognosis as well as tissue-specific methylation. *BMC Cancer* **11,** 380.

Newell-Price, J., Clark, A.J., King, P. (2000). DNA methylation and silencing of gene expression. *Trends in Endocrinology and Metabolism* **11(4),** 142-148.

Nold, M.F., Nold-Petry, C.A., Zepp, J.A., Palmer, B.E., Bufler, P., Dinarello, C.A. (2010). IL-37 is a fundamental inhibitor of innate immunity. *Nature Immunology* **11,** 1014-1022.

Nussbaumer, C., Gharehbaghi-Schnell, E., Korschineck, I. (2006). Messenger RNA profiling: a novel method for body fluid identification by real-time PCR. *Forensic Science International* **157(2-3),** 181-186.

Oakeley, E.J. (1999). DNA methylation analysis: A review of current methodologies. *Pharmacology & Therapeutics* **84(3),** 389-400.

Oehmichen, M., Zilles, K. (1984). Postmortem DNA and RNA synthesis. Preliminary studies in human cadavers. *Zeitschrift für Rechtsmedizin* **91(4),** 287-294.

Oh, J.J., Grosshans, D.R., Wong, S.G., Slamon, D.J. (1999). Identification of differentially expressed genes associated with HER-2/neu overexpression in human breast cancer cells. *Nucleic Acids Research* **27(20),** 4008-4017.

Ohtani, S., Yamamoto, T. (2010). Age estimation by amino acid racemization in human teeth. *Journal of Forensic Sciences* **55(6),** 1630-1633.

Ong, S.Y., Wain, A., Groombridge, L., Grimes, E. (2012). Forensic identification of urine using the DMAC test: a method validation study. *Science & Justice* **52(2),** 90-95.

Ou, X., Zhao, H., Sun, H., Yang, Z., Xie, B., Shi, Y., Wu, X. (2011). Detection and quantification of the age-related sjTREC decline in human peripheral blood. *International Journal of Legal Medicine* **125(4),** 603-608.

Pacheco, S.E., Houseman, E.A., Christensen, B.C., Marsit, C.J., Kelsey, K.T., Sigman, M., Boekelheide, K. (2011). Integrative DNA methylation and gene expression analyses identify DNA packaging and epigenetic regulatory genes associated with low motility sperm. *PLoS One* **6(6),** e20280.

Paliwal, A., Vaissiere, T., Herceg, Z. (2010). Quantitative detection of DNA methylation states in minute amounts of DNA from body fluids. *Methods* **52(3),** 242-247.

Park, S.M., Park, S.Y., Kim, J.H., Kang, T.W., Park, J.L., Woo, K.M., Kim, J.S., Lee, H.C., Kim, S.Y., Lee, S.H. (2013). Genome-wide mRNA profiling and multiplex quantitative RT-PCR for forensic body fluid identification. *Forensic Science International: Genetics* **7(1),** 143-150.

Parker, C., Hanson, E., Ballantyne, J. (2011). Optimization of dried stain co-extraction methods for efficient recovery of high quality DNA and RNA for forensic analysis. *Forensic Science International: Genetics Supplement Series* **3(1),** e309-e310.

Parson, W., Strobl, C., Huber, G., Zimmermann, B., Gomes, S.M., Souto, L., Fendt, L., Delport, R., Langit, R., Wootton, S., Lagacé, R., Irwin, J. (2013). Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). *Forensic Science International: Genetics* **7(5),** 543-549.

Partemi, S., Berne, P.M., Batlle, M., Berruezo, A., Mont, L., Riuró, H., Ortiz, J.T., Roig, E., Pascali, V.L., Brugada, R., Brugada, J., Oliva, A. (2010). Analysis of mRNA from human heart tissue and putative applications in forensic molecular pathology. *Forensic Science International* **203(1-3),** 99-105.

Patel, G., Peel, C. (2008). Identifying the origin of cells. *Forensic Science International: Genetics Supplement Series* **1(1),** 574-576.

Patel, J.L., Goyal, R.K. (2007). Applications of artificial neural networks in medical science. *Current Clinical Pharmacology* 2(**3),** 217-226.

Perez, C., Pascual, M., Martin-Subero, J.I., Bellosillo, B., Segura, V., Delabesse, E., Álvarez, S., Larrayoz, M.J., Rifón, J., Cigudosa, J.C., Besses, C., Calasanz, M.J., Cross, N.C., Prósper, F., Agirre, X. (2013). Aberrant DNA methylation profile of chronic and transformed classic Philadelphia-negative myeloproliferative neoplasms. *Haematologica* **98(9),** 1414-1420.

Phang, T.W., Shi, C.Y., Chia, J.N., Ong, C.N. (1994). Amplification of cDNA via RT-PCR using RNA extracted from postmortem tissues. *Journal of Forensic Sciences* **39(5),** 1275-1279.

Pilin, A., Pudil, F., Bencko, V. (2007). Changes in colour of different human tissues as a marker of age. *International Journal of Legal Medicine* **121(2),** 158-162.

Plachot, C., Lelievre, S.A. (2004). DNA methylation control of tissue polarity and cellular differentiation in the mammary epithelium. *Experimental Cell Research* **298(1),** 122-132.

Pogribny, I., Raiche, J., Slovack, M., Kovalchuk, O. (2004). Dose-dependence, sex- and tissue-specificity, and persistence of radiation-induced genomic DNA methylation changes. *Biochemical and Biophysica Research Communications* **320(4),** 1253-1261.

Polisecki, E.Y., Schreier, L.E., Ravioli, J., Corach, D. (2004). Common mitochondrial DNA deletion associated with sudden natural death in adults. *J ournal of Forensic Sciences* **49(6),** 1335-1338.

Price, E.M., Cotton, A.M., Lam, L.L., Farre, P., Emberly, E., Brown, C.J., Robinson, W.P., Kobor, M.S. (2013). Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics & Chromatin* **6(1),** 4.

Prinz, M., Tang, Y., Siegel, D., Yang, H., Zhou, B., Deng, H. (2011). Establishment of a fast and accurate proteomic method for body fluid cell type identification (U.S. Department of Justice, Ed.), pp. 1-61.

Promega (2013). MethylEdge Bisulfite Conversion System - Technical Manual.

Promega (2014). PowerPlex ESI 16 System - Technical Manual.

Psychogios, N., Hau, D.D., Peng, J., Guo, A.C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., Gautam, B., Young, N., Xia, J., Knox, C., Dong, E., Huang, P., Hollander, Z., Pedersen, T.L., Smith, S.R., Bamforth, F., Greiner, R., McManus, B., Newman, J.W., Goodfriend, T., Wishart, D.S. (2011). The human serum metabolome. *PLoS One* **6(2),** e16957.

Qiagen (2005). AllPrep DNARNA Mini Handbook.

Qiagen (2008). MinElute PCR Purification kit.

Qiagen (2009a). EpiTect Bisulfite Handbook.

Qiagen (2009b). EZ1 DNA Investigator Handbook.

Qiagen (2009c). PyroMark Gold Q96 Reagents Handbook.

Qiagen (2009d). PyroMark PCR Handbook.

Qiagen (2010a). Pyrosequencing - the synergy of sequencing and quantification.

Qiagen (2010b). QIAamp DNA Investigator Handbook.

Qiagen (2011). EZ1 DNA Blood Handbook.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13,** 341.

Raddatz, G., Hagemann, S., Aran, D., Sohle, J., Kulkarni, P.P., Kaderali, L., Hellman, A., Winnefeld, M., Lyko, F. (2013). Aging is associated with highly defined epigenetic changes in the human epidermis. *Epigenetics & Chromatin* **6(1),** 36.

Rakyan, V.K., Down, T.A., Maslau, S., Andrew, T., Yang, T.P., Beyan, H., Whittaker, P., McCann, O.T., Finer, S., Valdes, A.M., Leslie, R.D., Deloukas, P., Spector, T.D. (2010). Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Research* **20(4),** 434-439.

Rakyan, V.K., Down, T.A., Thorne, N.P., Flicek, P., Kulesha, E., Gräf, S., Tomazou, E.M., Bäckdahl, L., Johnson, N., Herberth, M., Howe, K.L., Jackson, D.K., Miretti, M.M,, Fiegler, H., Marioni, J.C., Birney, E., Hubbard, T.J., Carter, N.P., Tavaré, S., Beck, S. (2008). An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Research* **18(9),** 1518-1529.

Ramchandani, S., Bhattacharya, S.K., Cervoni, N., Szyf, M. (1999). DNA methylation is a reversible biological signal. *Proceedings of the National Academy of Sciences of the U.S.A.* **96(11),** 6107-6112.

Rando, O.J., Verstrepen, K.J. (2007). Timescales of genetic and epigenetic inheritance. *Cell* **128(4),** 655-668.

Reed, K., Poulin, M.L., Yan, L., Parissenti, A.M. (2010). Comparison of bisulfite sequencing PCR with pyrosequencing for measuring differences in DNA methylation. *Analytical Biochemistry* **397(1),** 96-106.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447,** 425-432.

Reik, W., Walter, J. (2001). Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics* **2(1),** 21-32.

Ribel-Madsen, R., Fraga, M.F., Jacobsen, S., Bork-Jensen, J., Lara, E., Calvanese, V., Fernandez, A.F., Friedrichsen, M., Vind, B.F., Højlund, K., Beck-Nielsen, H., Esteller, M., Vaag, A., Poulsen, P. (2012). Genome-wide analysis of DNA methylation differences in muscle and fat from monozygotic twins discordant for type 2 diabetes. *PLoS One* **7(12),** e51302.

Richard, M.L., Harper, K.A., Craig, R.L., Onorato, A.J., Robertson, J.M., Donfack, J. (2012). Evaluation of mRNA marker specificity for the identification of five human body fluids by capillary electrophoresis. *Forensic Science International: Genetics* **6(4),** 452-460.

Ritz-Timme, S., Laumeier, I., Collins, M. (2003). Age estimation based on aspartic acid racemization in elastin from the yellow ligaments. *International Journal of Legal Medicine* **117(2),** 96-101.

Rodwell, G.E.J., Sonu, R., Zahn, J.M., Lund, J., Wilhelmy, J., Wang, L., Xiao, W., Mindrinos, M., Crane, E., Segal, E., Myers, B.D., Brooks, J.D., Davis, R.W., Higgins, J., Owen, A.B., Kim, S.K. (2004). A transcriptional profile of aging in the human kidney. *PloS Biology* **2(12),** e427.

Roeder, A.D., Haas, C. (2013). mRNA profiling using a minimum of five mRNA markers per body fluid and a novel scoring method for body fluid identification. *International Journal of Legal Medicine* **127(4),** 707-721.

Ronaghi, M. (2000). Improved performance of pyrosequencing using single-stranded DNA-binding protein. *Analytical Biochemistry* **286(2),** 282-288.

Ronaghi, M. (2001). Pyrosequencing sheds light on DNA sequencing. *Genome Research* **11(1),** 3-11.

Rouge-Maillart, C., Jousset, N., Vielle, B., Gaudin, A., Telmon, N. (2007). Contribution of the study of acetabulum for the estimation of adult subjects. *Forensic Science International* **171(2-3),** 103-110.

Sabatini, L.M., Ota, T., Azen, E.A. (1990). Structure and sequence determination of the gene encoding human salivary statherin. *Gene* **89(2),** 245-251.

Sabatini, L.M., Ota, T., Azen, E.A. (1993). Nucleotide sequence analysis of the human salivary protein genes HIS1 and HIS2 and evolution of the STATH/HIS gene family. *Molecular Biology and Evolution* **10(3),** 497-511.

Sahin, K., Yilmaz, S., Temel, A., Gozukirmizi, N. (2011). DNA methylation analyses of monozygotic twins. *Abstracts/Current Opinion in Biotechnology* **22S,** S105.

Sakurada, K., Akutsu, T., Watanabe, K., Fujinami, Y., Yoshino, M. (2011). Expression of statherin mRNA and protein in nasal and vaginal secretions. *Legal Medicine (Tokyo)* **13(6),** 309-313.

Sakurada, K., Ikegaya, H., Fukushima, H., Akutsu, T., Watanabe, K., Yoshino, M. (2009). Evaluation of mRNA-based approach for identification of saliva and semen. *Legal Medicine (Tokyo)* **11(3),** 125-128.

Schutte, B., El Hajj, N., Kuhtz, J., Nanda, I., Gromoll, J., Hahn, T., Dittrich, M., Schorsch, M., Müller, T., Haaf, T. (2013). Broad DNA methylation changes of spermatogenesis, inflammation and immune response-related genes in a subgroup of sperm samples for assisted reproduction. *Andrology* **1(6),** 822-829.

Setzer, M., Juusola, J., Ballantyne, J. (2008). Recovery and stability of RNA in vaginal swabs and blood, semen, and saliva stains. *Journal of Forensic Sciences* **53(2),** 296-305.

Shen, C.J., Maniatis, T. (1980). Tissue-specific DNA methylation in a cluster of rabbit b-like globin genes. *Proceedings of the National Academy of Sciences of the U.S.A.* **77(11),** 6634-6638.

Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., Waterland, R.A., Issa, J.P. (2007). Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genetics* **3(10),** 2023-2036.

Sikirzhytskaya, A., Sikirzhytski, V., Lednev, I.K. (2012a). Raman spectroscopic signature of vaginal fluid and its potential application in forensic body fluid identification. *Forensic Science International* **216(1-3),** 44-48.

Sikirzhytskaya, A., Sikirzhytski, V., Lednev, I.K. (2012b). Raman spectroscopy coupled with advanced statistics for differentiating menstrual and peripheral blood. *Journal of Biophotonics* **7(1-2),** 59-67.

Singer-Sam, J., Tanguay, R.L., Rjggs, A.O. (1989). Use of Chelex to improve PCR signal from a small number of cells. *Amplifications: A Forum for PCR Users*, 11.

Small, K.S., Hedman, A.K., Grundberg, E., Nica, A.C., Thorleifsson, G., Kong, A., Thorsteindottir, U., Shin, S.Y., Richards, H.B., GIANT Consortium, MAGIC Investigators, DIAGRAM Consortium, Soranzo, N., Ahmadi, K.R., Lindgren, C.M., Stefansson, K., Dermitzakis, E.T., Deloukas, P., Spector, T.D., McCarthy, M.I., MuTHER Consortium. (2011). Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nature Genetics* **43(6),** 561-564.

Song, F., Luo, H., Li, Y., Zhang, J., Hou, Y. (2013). Characteristics of the two microbial markers in vaginal secretions in Chinese Han population. *Forensic Science International: Genetics Supplement Series* **4(1),** e312-e313.

Song, F., Mahmood, S., Ghosh, S., Liang, P., Smiraglia, D.J., Nagase, H., Held, W.A. (2009). Tissue specific differentially methylated regions (TDMR): Changes in DNA methylation during development. *Genomics* **93(2),** 130-139.

Song, Y.L., Kato, N., Liu, C.X., Matsumiya, Y., Kato, H., Watanabe, K. (2000). Rapid identification of 11 human intestinal Lactobacillus species by multiplex PCR assays using group- and species-specific primers dericed from 16S-23S rRNA intergenic spacer region and its flanking 23S rRNA. *FEMS Microbiology Letters* **187(2),** 167-173.

Spichenok, O., Budimlija, Z.M., Mitchell, A.A., Jenny, A., Kovacevic, L., Marjanovic, D., Caragine, T., Prinz, M., Wurmbach, E. (2011). Prediction of eye and skin color in diverse populations using seven SNPs. *Forensic Science International: Genetics* **5(5),** 472-478.

Steenhout, A., Pourtois, M. (1981). Lead accumulation in teeth as a function of age with different exposures. *British Journal of Indrustrial Medicine* **38(3),** 297-303.

Storm, T., Rath, S., Mohamed, S.A., Bruse, P., Kowald, A., Oehmichen, M., Meissner, C. (2002). Mitotic brain cells are just as prone to mitochondrial deletions as neurons: a large-scale single-cell PCR study of the human caudate nucleus. *Experimental Gerontology* **37(12),** 1389-1400.

Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z., Cedar, H. (2009). Developmental programming of CpG island methylation profiles in the human genome. *Nature Structural and Molecular Biology* **16(5),** 564-571.

Sumi, H., Naito, E., Dewa, K., Fukuda, M., Xu, H.D., Yamanouchi, H. (2005). Applicability of the parentally imprinted allele (PIA) typing of a VNTR upstream the H19 gene to forensic samples of different tissues. *Legal Medicine (Tokyo)* **7(3),** 179-182.

Suzuki, M.M., Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics* **9(6),** 465-476.

Suzuki, O., Oya, M., Katsumata, Y., Matsumoto, T., Yada, S. (1980). A new enzymatic method for the determination of spermine in human seminal stains. *Journal of Forensic Sciences* **25(1),** 99-102.

Svingen, T., Tonissen, K.F. (2006). Hox transcription factors and their elusive mammalian gene targets. *Heredity* **97(2),** 88-96.

Szyf, M. (2010). DNA methylation and demethylation probed by small molecules. *Biochimica et Biophysica Acta* **1799(10-12),** 750-759.

Takada, F., Vander Woude, D.L., Tong, H.Q., Thompson, T.G., Watkins, S.C., Kunkel, L.M., Beggs, A.H. (2001). Myozenin: an alpha-actinin- and gamma-filamin-binding protein of skeletal muscle Z lines. *Proceedings of the National Academy of Sciences of the U.S.A.* **98(4),** 1595-1600.

Tanaka, K., Okamoto, A. (2007). Degradation of DNA by bisulfite treatment. *Bioorganic and Medicinal Chemistry Letters* **17(7),** 1912-1915.

Teschendorff, A.E., Jones, A., Fiegl, H., Sargent, A., Zhuang, J.J., Kitchener, H.C., Widschwendter, M. (2012). Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Medicine* **4(3),** 24.

Teschendorff, A.E., Menon, U., Gentry-Maharaj, A., Ramus, S.J., Weisenberger, D.J., Shen, H., Campan, M., Noushmehr, H., Bell, C.G., Maxwell, A.P., Savage, D.A., Mueller-Holzner, E., Marth, C., Kocjan, G., Gayther, S.A., Jones, A., Beck, S., Wagner, W., Laird, P.W., Jacobs, I.J., Widschwendter, M. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research* **20(4),** 440-446.

Teschendorff, A.E., West, J., Beck, S. (2013). Age-associated epigenetic drift: implications, and a case of epigenetic thrift? *Human Molecular Genetics* **22(R1),** R7-R15.

Thompson, T.M., Sharfi, D., Lee, M., Yrigollen, C.M., Naumova, O.Y., Grigorenko, E.L. (2013). Comparison of whole-genome DNA methylation patterns in whole blood, saliva, and lymphoblastoid cell lines. *Behavior Genetics* **43(2),** 168-176.

Thorpe, S.R., Baynes, J.W. (1996). Role of the Maillard reaction in diabetes mellitus and diseases of aging. *Drugs & Aging* **9(2),** 69-77.

Tikhmyanova, N., Tulin, A.V., Roagiers, F., Golemis, E.A. (2010). Dcas supports cell polarization and cell-cell adhesion complexes in development. *PLoS One* **5(8),** e12369.

Trasler, J.M., Hake, L.E., Johnson, P.A., Alcivar, A.A., Millette, C.F., Hecht, N.B. (1990). DNA methylation and demethylation events during meiotic prophase in the mouse testis. *Molecular and Cellular Biology* **10(4),** 1828-1834.

Tsai, H.C., Baylin, S.B. (2011). Cancer epigenetics: linking basic biology to clinical medicine. *Cell Research* **21(3),** 502-517.

Tsuji, A., Ishiko, A., Takasaki, T., Ikeda, N. (2002). Estimating age of humans based on telomere shortening. *Forensic Science International* **126(3),** 197-199.

Tusnady, G.E., Simon, I., Varadi, A., Aranyi, T. (2005). BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. *Nucleic Acids Research* **33(1),** e9.

Vallone, P.M., Butler, J.M. (2004). AutoDimer: a screening tool for primer-dimer and hairpin structures. *BioTechniques* **37(2),** 226-231.

Valore, E.V., Park, C.H., Quayle, A.J., Wiles, K.R., McCray Jr, P.B., Ganz, T. (1998). Human beta-defensin-1: an antimicrobial peptide of urogenital tissues. *The Journal of Clinical Investigation* **101(8),** 1633-1642.

van den Berge, M., Carracedo, A., Gomes, I., Graham, E.A., Haas, C., Hjort, B., Hoff-Olsen, P., Maroñas, O., Mevåg, B., Morling, N., Niederstätter, H., Parson, W., Schneider, P.M., Syndercombe Court, D., Vidaki, A., Sijen, T. (2014). A collaborative European exercise on mRNA-based body fluid/skin typing and interpretation of DNA and RNA results. *Forensic Science International: Genetics* **10,** 40-48.

Vennemann, M., Koppelkamm, A. (2010). mRNA profiling in forensic genetics I: Possibilities and limitations. *Forensic Science International* **203(1-3),** 71-75.

Vidaki, A., Daniel, B., Syndercombe Court, D. (2013). Forensic DNA methylation profiling - Potential opportunities and challenges. *Forensic Science International: Genetics* **7(5),** 499-507.

Virkler, K., Lednev, I.K. (2009). Analysis of body fluids for forensic purposes: from laboratory testing to non-destructive rapid confirmatory identification at a crime scene. *Forensic Science International* **188(1-3),** 1-17.

Visser, M., Zubakov, D., Ballantyne, K.N., Kayser, M. (2011). mRNA-based skin identification for forensic applications. *International Journal of Legal Medicine* **125(2),** 253-263.

von Figura, G., Hartmann, D., Song, Z., Rudoplh, K.L. (2009). Role of telomere dysfunction in aging and its detection by biomarkers. *Journal of Molecular Medicine (Berlin, Germany)* **87(12),** 1165-1171.

von Zglinicki, T. (2000). Role of oxidative stress in telomere length regulation and replicative senescence. *Annals of New York Academy of Sciences* **908,** 99-110.

von Zglinicki, T., Martin-Ruiz, C.M. (2005). Telomeres as biomarkers for ageing and age-related diseases. *Current Molecular Medicine* **5(2),** 197-203.

Walsh, P.S., Metzger, D.A., Higuchi, R. (1991). Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *BioTechniques* **10(4),** 506-513.

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W., Kayser, M. (2013). The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Science International: Genetics* **7(1),** 98-115.

Warnecke, P.M., Stirzaker, C., Melki, J.R., Millar, D.S., Paul, C.L., Clark, S.J. (1997). Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Research* **25(21),** 4442-4426.

Warshauer, D.H., Lin, D., Hari, K., Jain, R., Davis, C., Larue, B., King, J.L., Budowle, B. (2013). STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic Science International: Genetics* **7(4),** 409-417.

Wasserstrom, A., Frumkin, D., Davidson, A., Shpitzen, M., Herman, Y., Gafny, R. (2013). Demonstration of DSI-semen--A novel DNA methylation-based forensic semen identification assay. *Forensic Science International: Genetics* **7(1),** 136-142.

Webb, J.L., Creamer, J.I., Quickenden, T.I. (2006). A comparison of the presumptive luminol test for blood with four non-chemiluminescent forensic techniques. *Luminescence* **21(4),** 214-220.

Weber-Lehmann, J., Schilling, E., Gradl, G., Richter, D. C., Wiehler, J., Rolf, B. (2014). Finding the needle in the haystack: differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing. *Forensic Science International: Genetics* **9,** 42-46.

Weidner, C.I., Lin, Q., Koch, C.M., Eisele, L., Beier, F., Ziegler, P., Bauerschlag, D.O., Jöckel, K.H., Erbel, R., Mühleisen, T.W., Zenke, M., Brümmendorf, T.H., Wagner, W. (2014). Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biology* **15(2),** R24.

Wild, L., Flanagan, J.M. (2010). Genome-wide hypomethylation in cancer may be a passive consequence of transformation. *Biochimica et Biophysica Acta* **1806(1),** 50-57.

Winnefeld, M., Brueckner, B., Gronniger, E., Stab, F., Wenck, H., Lyko, F. (2012). Stable ethnic variations in DNA methylation patterns of human skin. *The Journal of Investigative Dermatology* **132(2),** 466-468.

Wong, A.H., Gottesman, I., Petronis, A. (2005). Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Human Molecular Genetics* **14 (Spec No 1),** R11-R18.

Wong, E.M., Dobrovic, A. (2011). Assessing gene-specific methylation using HRM-based analysis. *Methods in Molecular Biology* **687,** 207-217.

Xiong, Z., Laird, P.W. (1997). COBRA: a sensitive and quantitative DNA methylation assay. *Nucleic Acids Research* **25(12),** 2532-2534.

Xu, H., Zhao, Y., Liu, Z., Zhu, W., Zhou, Y., Zhao, Z. (2012). Bisulfite genomic sequencing of DNA from dried blood spot microvolume samples. *Forensic Science International: Genetics* **6(3),** 306-309.

Xu, Y., Xie, J., Cao, Y., Zhou, H., Ping, Y., Chen, L., Gu, L., Hu, W., Bi, G., Ge, J., Chen, X., Zhao, Z. (2014). Development of highly sensitive and specific mRNA multiplex system (XCYR1) for forensic human body fluids and tissues identification. *PLoS One* **9(7),** e100123.

Xue, Y., Wang, Q., Long, Q., Ng, B.L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., Zhao, Y., Asan MacArthur, D.G., Quail, M.A., Carter, N.P., Yang, H., Tyler-Smith, C. (2009). Human Y Chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology* **19(17),** 1453-1457.

Yacubova, E., Komuro, H. (2002). Stage-specific control of neuronal migration by somatostatin. *Nature* **415(6867),** 77-81.

Yamada, N., Nishida, Y., Tsutsumida, H., Goto, M., Higashi, M., Nomoto, M., Yonezawa, S. (2009). Promoter CpG methylation in cancer cells contributes to the regulation of MUC4. *British Journal of Cancer* **100(2),** 344-351.

Yang, Y., Xie, B., Yan, J. (2014). Application of Next-generation Sequencing Technology in Forensic Science. *Genomics, Proteomics & Bioinformatics* **12(5),** 190-197.

Yarnell, A.T., Oh, S., Reinberg, D., Lippard, S.J. (2001). Interaction of FACT, SSRP1, and the high mobility group (HMG) domain of SSRP1 with DNA damaged by the anticancer drug cisplatin. *The Journal of Biological Chemistry* **276(28),** 25736-25741.

Yasunaga, M., Matsumura, Y. (2014). Role of SLC6A6 in promoting the survival and multidrug resistance of colorectal cancer. *Nature Scientific Reports* **4,** 4852.

Yi, H., Fujimura, Y., Ouchida, M., Prasad, D.D., Rao, V.N., Reddy, E.S. (1997). Inhibition of apoptosis by normal and aberrant Fli-1 and erg proteins involved in human solid tumors and leukemias. *Oncogene* **14(11),** 1259-1268.

Yi, S.H., Xu, L.C., Mei, K., Yang, R.Z., Huang, D.X. (2014). Isolation and identification of age-related DNA methylation markers for forensic age-prediction. *Forensic Science International: Genetics* **11,** 117-125.

Yu, Z., Zhai, G., Singmann, P., He, Y., Xu, T., Prehn, C., Römisch-Margl, W., Lattka, E., Gieger, C., Soranzo, N., Heinrich, J., Standl, M., Thiering, E., Mittelstraß, K., Wichmann, H.E., Peters, A., Suhre, K., Li, Y., Adamski, J., Spector, T.D., Illig, T., Wang-Sattler, R. (2012). Human serum metabolic profiles are age dependent. *Aging Cell* **11(6),** 960-967.

Zapico, S.C., Ubelaker, D.H. (2013). Applications of physiological bases of ageing to forensic sciences. Estimation of age-at-death. *Ageing Research Reviews* **12(2),** 605-617.

Zbieć-Piekarska, R., Spólnicka, M., Kupiec, T., Makowska, Ż., Spas, A., Parys-Proszek, A., Kucharczyk, K., Płoski, R., Branicki, W. (2015). Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. *Forensic Science International: Genetics* **14,** 161-167.

Zhang, Y., Wang, Z., Zhang, J., Farmer, B., Lim, S.H. (2008). Semenogelin I expression in myeloma cells can be upregulated pharmacologically. *Leukemia Research* **32(12),** 1889-1894.

Zhao, D., Ishikawa, T., Quan, L., Michiue, T., Zhu, B.L., Maeda, H. (2009). Postmortem quantitative mRNA analyses of death investigation in forensic pathology: an overview and prospects. *Legal Medicine (Tokyo)* **11 (Suppl 1),** S43-S45.

Zhao, G., Yang, Q., Huang, D., Yu, C., Yang, R., Chen, H., Mei, K. (2005). Study on the application of parent-of-origin specific DNA methylation markers to forensic genetics. *Forensic Science International* **154(2-3),** 122-127.

Zhu, B.L., Tanaka, S., Ishikawa, T., Zhao, D., Li, D.R., Michiue, T., Quan, L., Maeda, H. (2008). Forensic pathological investigation of myocardial hypoxia-inducible factor-1 alpha, erythropoietin and vascular endothelial growth factor in cardiac death. *Legal Medicine (Tokyo)* **10(1),** 11-19.

Zhuang, J., Jones, A., Lee, S.H., Ng, E., Fiegl, H., Zikan, M., Cibula, D., Sargent, A., Salvesen, H.B., Jacobs, I.J., Kitchener, H.C., Teschendorff, A.E., Widschwendter, M. (2012). The dynamics and prognostic potential of DNA methylation changes at stem cell gene loci in women's cancer. *PLoS Genetics* **8(2),** e1002517.

Zubakov, D., Hanekamp, E., Kokshoorn, M., van Ijcken, W., Kayser, M. (2008). Stable RNA markers for identification of blood and saliva stains revealed from whole genome expression analysis of time-wise degraded samples. *International Journal of Legal Medicine* **122(2),** 135-142.

Zubakov, D., Liu, F., van Zelm, M.C., Vermeulen, J., Oostra, B.A., van Duijin, C.M., Driessen, G.J., van Dongen, J.J., Kayser, M., Langerak, A.W. (2010). Estimating human age from T-cell DNA rearrangements. *Current Biology* **20(22),** R970-R971.

Zykovich, A., Hubbard, A., Flynn, J.M., Tarnopolsky, M., Fraga, M.F., Kerksick, C., Ogborn, D., MacNeil, L., Mooney, S.D., Melov, S. (2014). Genome-wide DNA methylation changes with age in disease-free human skeletal muscle. *Aging Cell* **13(2),** 360-366.

ZymoResearch (2013a). EZ DNA Methylation Kit - Instruction Manual.

ZymoResearch (2013b). ZymoTaq PreMix.

Appendix I. Research Ethics documents

Approval letter

Miss Athina Vidaki
Department of Forensic and Analytical Science
Room 4.124 Franklin-Wilkins Building
150 Stamford Street
London
SE1 9NH

27 January 2014

Dear Athina,

BDM/13/14-30 DNA methylation profiling for various forensic applications.

Review Outcome: Full Approval

Thank you for sending in the amendments/clarifications requested to the above project. I am pleased to inform you that these meet the requirements of the BDM RESC and therefore that full approval is now granted.

Please ensure that you follow all relevant guidance as laid out in the King's College London Guidelines on Good Practice in Academic Research (http://www.kcl.ac.uk/college/policyzone/index.php?id=247).

For your information ethical approval is granted until 27 January 2017. If you need approval beyond this point you will need to apply for an extension to approval at least two weeks prior to this explaining why the extension is needed, (please note however that a full re-application will not be necessary unless the protocol has changed). You should also note that if your approval is for one year, you will not be sent a reminder when it is due to lapse.

Ethical approval is required to cover the duration of the research study, up to the conclusion of the research. The conclusion of the research is defined as the final date or event detailed in the study description section of your approved application form (usually the end of data collection when all work with human participants will have been completed), not the completion of data analysis or publication of the results. For projects that only involve the further analysis of pre-existing data, approval must cover any period during which the researcher will be accessing or evaluating individual sensitive and/or un-anonymised records. Note that after the point at which ethical approval for your study is no longer required due to the study being complete (as per the above definitions), you will still need to ensure all research data/records management and storage procedures agreed to as part of your application are adhered to and carried out accordingly.

If you do not start the project within three months of this letter please contact the Research Ethics Office.

Should you wish to make a modification to the project or request an extension to approval you will need approval for this and should follow the guidance relating to modifying approved applications:
http://www.kcl.ac.uk/innovation/research/support/ethics/applications/modifications.aspx
The circumstances where modification requests are required include the addition/removal of participant groups, additions/removal/changes to research methods, asking for additional data from participants, extensions to the ethical approval period. Any proposed modifications should only be carried out once full approval for the modification request has been granted.

Any unforeseen ethical problems arising during the course of the project should be reported to the approving committee/panel. In the event of an untoward event or an adverse reaction a full report must be made to the Chair of the approving committee/review panel within one week of the incident.

Please would you also note that we may, for the purposes of audit, contact you from time to time to ascertain the status of your research.

If you have any query about any aspect of this ethical approval, please contact your panel/committee administrator in the first instance (http://www.kcl.ac.uk/innovation/research/support/ethics/contact.aspx). We wish you every success with this work.

Yours sincerely,
James Patterson – Senior Research Ethics Officer

Cc: Denise Syndercombe Court

INFORMATION SHEET FOR PARTICIPANTS

This study has been approved by King's College London Biomedical Sciences, Dentistry, Medicine and Natural and Mathematical Sciences Research Ethics Subcommittee, BDM/13/14-30

DNA methylation profiling for various forensic applications

We would like to invite you to participate in postgraduate research. You should only participate if you want to; choosing not to take part will not disadvantage you in any way. Before you decide whether you want to take part, it is important for you to understand why the research is being done and what your participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information.

The Research

I am a PhD student enrolled in the Department of Forensic and Analytical Science at King's College London, completing the degree of PhD in Forensic Genetics. For completion, we are carrying out research projects sponsored by King's College London, with funding from the European Forensic Genetics Network of Excellence (EuroForGen) or DNA Analysis at King's.

The research involves an investigation into the ability to use DNA in various forensic applications, such as identifying the tissue source of a stain or the biological age of an individual. Body fluid samples will undergo DNA methylation profiling as well as DNA profiling will be carried out. In more detail, possible DNA modifications of several locations within the genome will be investigated in order to identify any meaningful variations.

These studies will develop methods to improve the way that evidence is analysed and interpreted in forensic and police investigations. You are invited to assist in this research by volunteering biological samples, and we would be very appreciative of your donation to our work. Volunteers will be asked to provide blood, saliva, sweat, semen, urine samples as well as vaginal fluid, menstrual fluid, cheek, skin or nasal fluid swabs. (NOTE: you are able to donate all samples required if this is possible to do so, but you can donate as many or as few samples as you wish).

You are invited to assist in this research by volunteering biological samples if you are over the age of 18, have no recent history of skin conditions or adverse reactions to blood donation, and we would be very appreciative of your donation to our work. For volunteers under the age of 18, the mother is asked to consent on behalf of the child.

Should you wish to take part in this study, you will be asked to make any of the following small body fluid donations. It is up to you which samples you wish to donate for the research.

- **Blood:** Blood samples will only need to be taken once. Please let us know if you have ever had an adverse reaction to donating blood. A 20 ml sample (less than 4 teaspoons) will be taken from your arm by a qualified and experienced medical professional. Blood collection will take place at a Study Suite, Room FWB 4.25.
- **Saliva, Sweat, Semen and Urine:** You will be provided with receptacles for saliva, sweat, semen and urine donations, which you will be asked to collect in the privacy of your home. Please keep the samples refrigerated whilst at home.
- **Cheek, Skin, Vaginal Fluid, Menstrual Fluid and Nasal Fluid Swabs:** You will be given a swab and asked to gently rub the inside of your cheek, forearm or vagina or nose.

Volunteering for this study will involve sample donation only, and should not take longer than 30 minutes of your time. There will be a minor risk of bruising during collection of blood. To minimise this risk, an experienced phlebotomist will perform blood collection, which will also take place in the presence of a trained first aider. There will also be a minor risk of injury due to abrasion during mouth swab collection and fingerprint deposition.

DNA will be extracted from these samples for the purpose of DNA methylation profiling.

DNA methylation profiling is a technique by which differences in peoples' DNA modifications are used to identify the source of the biological fluid or estimate the age of an individual.

If you would like more information on the process, please feel free to ask!

No further genetic information will be assessed on your samples and you will receive no feedback on the analysis of your DNA. Your DNA samples will be stored until 31st December 2016, when the EuroForGen project comes to an end. All the data/records taken will be stored in a secure location in King's College London and stored for a period of up to 2 years after the end of the EuroForGen project to allow for publications, so until 31st December 2018. These data/records will only be accessed by me (Athina Vidaki) and my supervisor (Dr Denise Syndercombe Court) and will also be shared with our colleagues (EuroForGen, KCL).

A DNA profile can, in theory, be used to provide personal information about an individual, including susceptibility to medical conditions and inheritable disease. However, we will not be carrying out any analysis that would show susceptibility to any medical conditions. To avoid concern over your privacy, the source of each donation will be anonymised and confidential. Each sample is given an individual code number so that DNA profiles will not be able to be traced back to their original donor.

The DNA methylation profiling results will be published at the end of the research in the form of a PhD thesis and scientific publications, with the original donor being anonymised. You will be allowed to look through the report once it is complete, but no individual feedback on test results will be shared with you as the investigator will not know the identity of the original donor.

At any time you may choose to withdraw your consent. Should you chose to do so, no explanation is needed and all your samples and any related data will be removed and destroyed. Please do so before 1st March 2014.

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep and be asked to sign a consent form. If you decide to take part you are still free to withdraw at any time before 1st March 2014 and without giving a reason.

For further information on the study please contact:

Miss Athina Vidaki, Department of Forensic and Analytical Science, Room 4.124, 4th Floor Franklin Wilkins Building, King's College London, 150 Stamford Street, London, SE1 9NH.
Email: athina.a.vidaki@kcl.ac.uk

If this study has harmed you in any way you can contact King's College London using the contact details below for further advice and information.

Dr Denise Syndercombe Court, Department of Forensic and Analytical Science, Room 4.109, 4th Floor Franklin Wilkins Building, King's College London, 150 Stamford Street, London, SE1 9NH.
Email: denise.syndercombe-court@kcl.ac.uk

CONSENT FORM FOR PARTICIPANTS IN RESEARCH STUDIES

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

KING'S College LONDON
University of London

Title of Study: DNA methylation profiling for various forensic applications

King's College Research Ethics Committee Ref: BDM/13/14-30

Thank you for considering taking part in this research. The person organizing the research must explain the project to you before you agree to take part. If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.

- ***I understand that if I decide at any other time during the research that I no longer wish to participate in this project, I can notify the researchers involved before 1st March 2014 and be withdrawn from it immediately (please tick if you agree)*** ☐

- ***I consent to donating the following samples for DNA methylation profiling (tick those that apply)***

  ☐ *Saliva*
  ☐ *Blood*
  ☐ *Cheek Swab*
  ☐ *Semen*
  ☐ *Urine*
  ☐ *Sweat*
  ☐ *Vaginal Fluid Swab*
  ☐ *Skin Swab*
  ☐ *Menstrual Fluid Swab*
  ☐ *Nasal Fluid Swab*

- ***I understand that I will not receive any results about my own DNA profile (please tick if you agree)*** ☐

- ***I consent to the processing of my personal information for the purposes of this research study. I understand that such information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 1998 (please tick if you agree)*** ☐

- ***I agree that the research team may use my data for future research and understand that any such use of identifiable data would be reviewed and approved by a research committee. (In such cases, as with this project, data would not be identifiable in any report) (please tick if you agree)*** ☐

Participant's Statement:

I _____

agree that the research project named above has been explained to me to my satisfaction and I agree to take part in the study. I have read both the notes written above and the Information Sheet about the project, and understand what the research study involves.

Signed **Date**

Investigator's Statement:

I _____

confirm that I have carefully explained the nature, demands and any foreseeable risks (where applicable) of the proposed research to the volunteer.

Signed **Date**

Appendix II. Bisulphite Pyrosequencing® assay design

Example: **BLood Methylated marker 1 – BLM1**

Gene: Solute Carrier Family 6, member 6 (SLC6A6)

CpG Position: Chr3: 14,443,428

Sequence used in primer design: Chr3 : 14,443,256–14,443,605 (350 bp)

Sense strand:

```
 5′-TTCGGGGTTCTCACAGCGTTTCCCTCTCATGCTGCCAGGGCGGGGCAGTGGAAAATGGTCCCTGAG
GGCCCGTTCATTCAAATAAGGACAAGAATAATGGCACACTCCCCATCTCCATTTAAAAATAGTCACATTA
GCTGACACTGGCTTGGCACTTCCTAGATGCCGGGTGCGTTCCGAGTTCTCTACGCTGTTAACCTTGGATC
TTCATAACAGCCGGAGGAGACAAGCGCTACTATTTTCCTTTCCATTTTATAGATGAGGACACTGAGGCAG
AGAAATTGAGTCACTTGCCCAGGGTCACACAGCCATGGCAAGGCCAGGATTCTGGAGCTCTGTACCTGGC
TTCA-3′
```

DNA Sequence after bisulphite treatment:

```
    1 TTCGGGGTTCTCACAGCGTTTCCCTCTCATGCTGCCAGGGCGGGGCAGTGGAAAATGGTC
      ||++||||||:|:|:||++|||:::|:|:|||:||::|||||++|||:|||||||||||||:
    1 TTCGGGGTTTTTATAGCGTTTTTTTTTTATGTTGTTAGGGCGGGGTAGTGGAAAATGGTT

   61 CCTGAGGGCCCGTTCATTCAAATAAGGACAAGAATAATGGCACACTCCCCATCTCCATTT
      ::|||||||::++||:|||:|||||||||:||||||||||||:|:|:|::::||:|::||||
   61 TTTGAGGGTTCGTTTATTTAAATAAGGATAAGAATAATGGTATATTTTTTATTTTTATTT

  121 AAAAATAGTCACATTAGCTGACACTGGCTTGGCACTTCCTAGATGCCGGGTGCGTTCCGA
      |||||||||:|:||||||:|||:|:|||:||||:|:|||::|||||||:++||||++||:++|
  121 AAAAATAGTTATATTAGTTGATATTGGTTTGGTATTTTTTAGATGTCGGGTGCGTTTCGA

  181 GTTCTCTACGCTGTTAACCTTGGATCTTCATAACAGCCGGAGGAGACAAGCGCTACTATT
      |||:|:||++:|||||||::|||||||:||:||||:||:++|||||||:|||++:||:||||
  181 GTTTTTTACGTTGTTAATTTTGGATTTTTATAATAGTCGGAGGAGATAAGCGTTATTATT

  241 TTCCTTTCCATTTTATAGATGAGGACACTGAGGCAGAGAAATTGAGTCACTTGCCCAGGG
      ||::|||::|||||||||||||||||:|:||||||:|||||||||||||||:|:|||:::||||
  241 TTTTTTTTTATTTTATAGATGAGGATATTGAGGTAGAGAAATTGAGTTATTTGTTTAGGG

  301 TCACACAGCCATGGCAAGGCCAGGATTCTGGAGCTCTGTACCTGGCTTCA
      |:|:|:||::||||:||||::|||||||:|||||:|:||||::||||:||:|
  301 TTATATAGTTATGGTAAGGTTAGGATTTTGGAGTTTTGTATTTGGTTTTA


(Assuming all CpG sites are methylated)
```

Primer design parameters (screenshot from BiSearch software):

Sequence	TTCGGGGTTCTCACAGCGTTTCCCTCTCATGCTGCCAGG
Bisulphite	☑ Max Tm difference = 2.5°C
Set search region for	Forward primer: - Reverse primer: -
Max length of PCR	200

Primer melting temperature

Primer conc	1.0 mikromol	Glycerol conc	0.0 %
Potassium conc	50.(milimol	Ethylen glycol conc	0.0 %
Magnesium conc	1.5 milimol	Formamidconc	0.0 %

Primer scoring values

Description	Weight	Min	Opt	Max	Description	Weight	Max
Primer length	0.5	18	23	30	Self annealing	0.1	20
GC content	1.0	40.(	50.(	60.(	Self end-annealing	0.2	10
GC content (bis)	1.0	0.0	30.(	60.(	Pair annealing	0.1	20
Melting temp	1.0	45.(	55.(	65.(	Pair end-annealing	0.2	10

Primer design

Results list size	10	Max Tm diff	3.0	Minimum of CpGs	1

Database search and fast PCR

Database	Homo sapiens	Mismatches	000000001 Max: 1
PCR length	1000	PCR product to show	100
		Primer matches to show	100

Primer design results:

No	Sc	Primer sequences	Pos	Plen	%GC	Tm	oTm	CpG	Sa	Sea	Pa	Pea	Len
1.	11.11	TAGTTGATATTGGTTTGGTA	135	20	30.0	55.3	55.3	6	12	4	18	4	159
		CAAATAACTCAATTTCTCTAC	293	21	28.6	54.7	54.7		20	0	Details		

Abbreviations used:

Sc: Score
Fpcr: Do Fast (electronic) PCR with the primer pairs
Pseq: Primer sequence
Pos: Position of the primer
Plen: Primer length
%GC: Content of G and C in percent
Tm: Melting temperature
oTm: Melting temperature of original primer (C are not translated to T)
CpG: Number of CpG islands in the unmodified PCR product
Sa: Self annealing
Sea: Self end annealing
Pa: Pair annealing
Pea: Pair end-annealing
Len: Length of the PCR product

Primers:

Forward primer: 5′- TAG**T**TGA**T**A**T**TGG**T**TTGG**T**A – 3′

Reverse primer: 5′- CAA**A**T**A**ACTCAATTTCTCT**A**C -3′

(Converted cytosines in bold)

Possible primer dimers:

The best self annealing of the forward primer:

```
    TAGTTGATATTGGTTTGGTA
     |....||||....|
ATGGTTTGGTTATAGTTGAT
```

The best self end-annealing of the forward primer:

```
TAGTTGATATTGGTTTGGTA
       ||.........||
       ATGGTTTGGTTATAGTTGAT
```

The best self annealing of the reverse primer:

```
 CAAATAACTCAATTTCTCTAC
  |.|.||.|..|.||.|.|
CATCTCTTTAACTCAATAAAC
```

The best self end-annealing of the reverse primer: None

The best pair annealing of these primers:

```
TAGTTGATATTGGTTTGGTA
  ||.||.|
  CATCTCTTTAACTCAATAAAC
```

The best pair end-annealing of these primers:

```
TAGTTGATATTGGTTTGGTA
                   |
               CATCTCTTTAACTCAATAAAC
```

Wrong annealing of the forward primer with the PCR product:

```
5′- ATTATAAAAATCCAAAATTAACAACRTAAAAAACTCRAAACRCACCCRACATCTAAA -3′
          ||..|.....||.||
 3'-ATGGTTTGGTTATAGTTGAT-5'
```

Possible hairpin formation:

The best hairpin forming of the forward primer:

```
 TGAT-5'
T  ||
 GATATTGGTTTGGTA-3'
```

The best hairpin forming of the reverse primer:

```
 CAATAAAC-5'
{   ||||
 TCAATTTCTCTAC-3'
```

The best self end-annealing of the PCR product:

```
 ACTA-3'
A  ||
CTATAACCAAACCATAAAAAATCTACANCCCACNCAAANCTCAAAAAATNCAACAATTAAAACCTAAAAA
TATTATCANCCTCCTCTATTCNCAATAATAAAAAAAAAAATAAAATATCTACTCCTATAACTCCATCTCT
TTAACTCAATAAAC-5'
```

Database:

PCR product(s) on the bisulphite transformed sense chain:

```
      Forward primer: TAGTTGATATTGGTTTGGTA

      Reverse primer: CAAATAACTCAATTTCTCTAC


1. Chromosome 3 (len: 159) (Designed product)


          14443389                        TAGTTGATATTGGTTTGGTATT
          TTTTAGATGT TGGGTGTGTT TTGAGTTTTT TATGTTGTTA ATTTTGGATT
          TTTATAATAG TTGGAGGAGA TAAGTGTTAT TATTTTTTTT TTTATTTTAT
          AGATGAGGAT ATTGAGGTAGAGAAATTGAGTTATTTG      14443548
```

Matches of forward primer: 854 matches based on the 3' 16mer oligo search.

Matches of reverse primer: 601 matches based on the 3' 16mer oligo search.

PCR product(s) on the bisulphite transformed antisense chain:

```
      Forward primer: TAGTTGATATTGGTTTGGTA

      Reverse primer: CAAATAACTCAATTTCTCTAC


1. Chromosome 2 (len: 728) (Non-specific product)


 185716536    ACAAAAACTCCATTTCTCTACATACTTA TCAACACTTA CTATCTTTTA
              TATTTTACTA ACAAACATCC TAACAAATTT AAAATAATAT CTCATCATAA
              TTTTACTTTA CATTTCTCTA ATAATTAATA ATACTAAACA TTTTATTTCA
              TATATCTTTA AAACATTCAT ATATCTTCTT CTAAAAAATA TCTATTCAAA
              TCCTTTACCC ATTTTTAATC AAATTATTTA TTTTCTTAAT CTTTAATCAT
              TTAAATTTCT TATCAATTTT AATTATTAAA CCTTATCAAA TTTATAATTA
              AAAATACAAT AATCACCTAC TTATCTATAA AAAATATTTT CCAAAATCTA
              CAATAAATAC TTAAAATCAC AAATAATACT AATAATATAC TAAATATAAA
              AATTATATAA ATACAATCTC TTTCTCTCTC TCTCTCTCTC TCTCTCTCTC
              TCTCTATCTC TCAAAATATC CAATTATACT ACATTCACCT ATTTTTTAAA
              TAATTAACCA TCAATAACTA AAACTACTAA AAACAAAACT ATAAATAAAA
              AAAAACTACT ATATTTTTCC TATCCTATTA ATTATCTCTT CATCCTATTA
              ATTATTTCCT TAACTATACA AAACTTTATA CTTTAAAATT ATCTCATTTA
              ACTATACTTT TTTTTTTTTT TTTTTAACCT ATACTTTTAA AATCACTACC
              AAAAATATAT TACCAAACTAATATCATAAA                185717264
```

Matches of forward primer: 796 matches based on the 3' 16mer oligo search.

Matches of reverse primer: 621 matches based on the 3' 16mer oligo search.

PCR product DNA double-stranded sequence:

Sequencing Primer

Forward Primer

```
5′- TAGTTG ATATTGGTTT GGTATTTTTT AGATGTYGGG TGYGTTTYGA GTTTTTTAYG
3′- ATCAAC TATAACCAAA CCATAAAAAA TCTACARCCC ACRCAAARCT CAAAAAATRC

TTGTTAATTT TGGATTTTTA TAATAGTYGG AGGAGATAAG YGTTATTATT TTTTTTTTTA
AACAATTAAA ACCTAAAAAT ATTATCARCC TCCTCTATTC RCAATAATAA AAAAAAAAAT

TTTTATAGAT GAGGATATTG AGGTAGAGAA ATTGAGTTAT TTG -3′
AAAATATCTA CTCCTATAAC TCCATCTCTT TAACTCAATA AAC -5′
```

Reverse Primer - Biotin

Y: C or T
R: G or A

Amplicon Length: 159 bp

Template Loop: No (Last 6 bp at 3'-end → GTTGAT)

Sequencing Primer:

5′- A**T**A**T**TGG**T**TTGG**T**A**T**TT**TT**TAGAT – 3′ (Converted cytosines in bold)

Tm: 58.0 °C (BiSearch)

Wrong annealing of the sequencing primer with the PCR-product:

```
5'-CAAATAACTCAATTTCTCTACCTCAATATCCTCATCTATAAAATAAAAAAAAAAATAATAACRCTT
                                    |||||x||||||xx|||xx||xxx
                                 3'-TAGATTTTTTATGGTTTGGTTATA-5'
```

DNA methylation assay:

Sequence to analysed: GC**C**GGGTG**C**GTTC**C**GAGTTCTCTA**C**GCT

Sequence to analysed bisulphite-treated: G**T**YGGGTGYGTT**T**YGAGTT**T**T**T**TAYG**T**T

Number of CpG sites: 4

Dispensation order: AGTCGTGATCGTCGAGTCAGTCGT

Chromosomal Location: Chr3: 14,443,420 – 14,443,469

Sequenced strand: Sense

Biotin modification: Reverse PCR primer

Appendix III. MiSeq® auto-analysis command script for age markers

```
for sample in *_L001_R1_001.fastq.gz
# Get the name of the sample from the file names and designate $describer 2 as this
name
do
echo $sample
describer=$(echo ${sample} | sed 's/_L001_R1_001.fastq.gz//')
echo $describer
describer2=$(echo $describer | rev | cut -c 5- | rev)
echo $describer2
# Create sam files aligned to 16AgeMarkers ref
/users/davidballard/bwa/bwa mem
/users/davidballard/bwa/References/16AgeMarkers.fa
${describer}_L001_R1_001.fastq.gz ${describer}_L001_R2_001.fastq.gz >
$describer2.sam
done
# Convert file from SAM to BAM format
/users/davidballard/samtools/samtools view -bt
/users/davidballard/samtools/References/16AgeMarkers.fa.fai $describer2.sam >
${describer2}.uns.bam
# Sort BAM file
/users/davidballard/samtools/samtools sort ${describer2}.uns.bam ${describer}
# Index the bam file
/users/davidballard/samtools/samtools index ${describer}.bam
# Remove intermediate files
rm ${describer2}.uns.bam
# Remove sam files
rm $describer2.sam
# Create read groups
java -jar /Users/davidballard/GATK/AddOrReplaceReadGroups.jar I=
$describer.bam O= $describer2.bam SORT_ORDER=coordinate RGID=$describer2
RGLB=bar RGPL=illumina RGPU=run RGSM=$describer2
CREATE_INDEX=True
# Remove intermediate files
rm ${describer}.bam
rm ${describer}.bam.bai
# Create variant calls for all age SNPs
java -jar /Users/davidballard/GATK/GenomeAnalysisTK.jar -T UnifiedGenotyper -
R /Users/davidballard/GATK/16AgeMarkers.fa -I $describer2.bam -glm SNP --
alleles /Users/davidballard/GATK/Age.vcf --genotyping_mode
GENOTYPE_GIVEN_ALLELES -stand_emit_conf 10 -stand_call_conf 30 -o
$describer2.vcf --output_mode EMIT_ALL_SITES --downsampling_type none --
dbsnp /Users/davidballard/GATK/Age.vcf
done
```

Appendix IV. Expected bisulphite-converted DNA sequences of designed PCR products (16 age CpG markers in yellow)

> cg19761273
TGTTTAGTTTGAAGATTGAGGATAAAGTTATTATTTTTTATAAAATGAGGTTAGATT
ATTTGTTTTTTTTTTAGTTTTTGCGGTTTGGAGACGGAGTTAATATTTTTATTTGTGT
TGGATTTGAATGTTTTTTTTTTGTAAAGGAAATAAGG

> cg27544190
GGGTAGGATTAAAGTTGAGGCGTGTTTATAGATATTTGTTTGGTGTGAGTTTTTGGT
ATATAGATGGTTGCGAGTGAAGTGGCGACGGGGGATTGTTATTTTTAAG

> cg03286783
GTTTTAGTTAGTGGGTGGGGTTAGGGTATATTCGTTTTTTTTTCGGTTTTTTTTTTCG
TTTTTTAATCGTGAGGTGTTGGGTTTGGGGGACGTTGGTAGTTGGGTTTTTTCGGTT
TTTTTGGGTAGGTGTAGGGTCGGGTTTAAAGTTTTCGGAACGCGTTTTGGTTTGATT
TGAGGAGGGG

> cg01511567
TATTAGATTTAGTATAGGGGTGGGGTGGGGGTGTGTATTGGAATGATGCGTGTTCG
TTTTTTTGTAAAATAGTTTTTATGTTATGGAAAGGAGTCGATGGGATAAGAAGAGAA
GTATTTGAATAGTTGTGGG

> cg07158339
GGAATATGTTTTGTTTAAAAAAAATTTTATAGGGTTTAATTTATTTTATTTTTATTAT
AATTTTATGAAGTAGGAATTTTTATAAAACGTATTTTATAAATAAGGTATAGAGAGG
TTAATTA

> cg05442902
GTATGTTTTGGTTTTTTGTATACGTTGTTTTTTTGTATTAGGAATTTTTATTTTTATTTT
TGTTTGTTTGTCGAATTTTAGAAATTTGTAAGGGTTAGTTTAGAGGTTATT

> cg24450312
GTTATTTATAGAGTTTGAGCGGTTTCGAGTTATCGTCGTCGTTTTTCGATCGGTTTT
CGCGTTTTTGTGTTTCGGCGTTTTTTTTTATTGCGTTCGGGGCGCGCGAGGGGCGTA
GCGTTCGGAGGGTTGTTCGGGGGAATTTGGAGTTTTCGTTTCGGGTTTTTCGATTC
GTTCGTTCGTTTCGGTTTGGTTTGTAGTAGA

> cg17274064
AGGGAATAAGTATTTTTTTAATTTGAAAAATAATAATTAAAATTTTTTTAATTTTTAA
GGTCGAGTAATATAATTTATTAATTGGTCGTTATTAATATGTAGTTTTATTGATTAT
AGTATATAGAAGTTTGATTGTGAG

> cg02085507
GTTAATGGATTTGGTTTTGGCGAAGGCGCGTTTTGGGTTGGATCGAAATTTTTTTAT
TCGTTTTGTGGTCGGAGGGATTAGATTATTAGTGGGACGAATTTAGACGTTTGGAG
TCGGGTCGGTTCGTAAGAACGGGGTGATTTTAGGTTGTTTTTGTAATGGGGTTGAG
GAAGGATTTTTTGAGTT

> cg20692569
TTGTTGTTGTGGTAGTTGTTGGCGGCGGGCGGCGCGGTATTGGAGATCGGTCGTTT
CGATTCGGAGCGCGGGCGCGGGGTTGCGTCGTGTTAGGCGGTGGAGATTTTTATGT
GTCGCGGTATCGGTTATAATTTGATTCGTATGTTTAATTTGTTGGGTT

> cg04528819
AATAGGTTTTGGTGTAGTTCGGGAAGGGGTATTGGTGGCGTTTGGTAGTAGGTGTG
ATAGATTTTTTTCGGGGCGTTTGATTCGCGGCGGGGGCGGGGTTTGTTTTTAGGGT
TTTTTTAGAGAATTTATTAGAGGTTG

> cg08370996
GTGTTAAAGTTTATTATATAGAGAGTTTAGTGAGTTGATCGCGGAGAAGTTATTTTT
GTTAGTTTCGGCGTTTATAAATCGTATTTTTTTTCGCGTTTTTTTTTTTAGTATATTT
GATTATTTTGATTTTTTGTTTTTTTTTTTTCGCGGTGTGTGTGCGTGCGCGCGTG
TGTGTTTTTTTTTTT

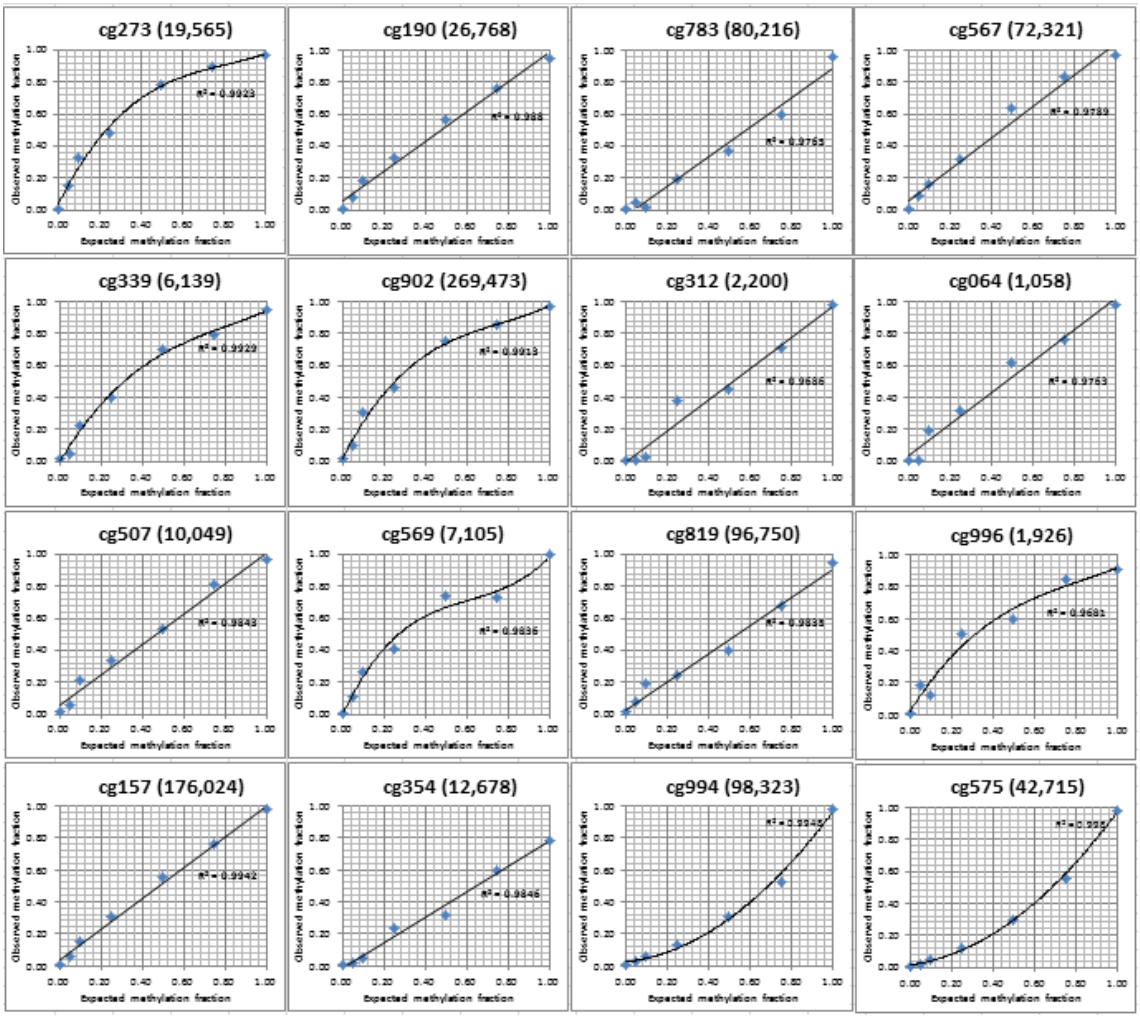
> cg04084157
GAGGGTGTTTGTTTTTTTCGGTTTGCGTTTGCGCGTTGGGGTTTTCGGTTGAAGGG
GTGTGCGTTAGCGGAGTTTCGGGAAATGAATGAATGAATGAATGAATGAAATGTT

> cg22736354
GTTGAGTTTAGGAGTTTTATGAGGTGTAGTATCGGTAGGTAGTCGTTGGTGTCGTA
GTTTCGGTAAGTTCGTTTGTAGAATGGGTATTCGAGGGTTAGAGTGCGCGGGTGCG
TTAGGGCGGTTACGTAGGTTAGGTAGATTACGTGGTCGTAGGATAGGTTGCGCGGG
CGTCGTTGTTGTCGGTGGTTAAATTTTTTAAAG

> cg06493994
GGAGAGTAAGTTAAGAAATACGGTGAAGGAGTTTTTTTTAAAGTTGTTTAGGTTTTT
TCGCGTCGGTGTTTGG
TTTTCGTCGTTAATATTATGGATAGTTTTCGGGAATCGATTTTGGGGCGTTTGGACG
TCGTTGGTTTTTGGTAGGTT

> cg02479575
GGAGGAGAATGTTATTTATTTTAGTATTAAATATTCGGATAGCGTTTTTCGGGAGGT
TCGAGAAGAGAATCGCGATTTGTTTTAGTATCGGGGTTTAGGATAGTTTTTAGCGGG
TTTCGTTTCGTTTTTAGAATTTTGGATAG

Appendix V. Expected *vs.* observed methylation of methylation controls for all 16 age-associated CpGs (MiSeq®)



The graphs above represent the individual 'normalised' observed vs. expected methylation for all 16 age-associated CpG sites as quantified in predefined DNA methylation standards (0%-100%) on MiSeq® using 30 PCR cycles. Each CpG is identified by the last three numbers of its name (for example cg273 corresponds to cg19761273), while numbers in brackets correspond to average read numbers. As shown for most markers, graphs fit linear regression lines (e.g.cg567), while for some others significant amplification bias towards either the methylated (e.g. cg273) or the unmethylated (e.g. cg994) were observed.

Appendix VI. List of associated publications

Original Articles in Peer-Reviewed Scientific Journals

C. Haas, E. Hanson, M.J. Anjos, R. Banemann, A. Berti, E. Borges, A. Carracedo, M. Carvalho, C. Courts, G. De Cock, M. Dotsch, S. Flynn, I. Gomes, C. Hollard, B. Hjort, P. Hoff-Oslen, K. Hribikova, A. Lindenbergh, B. Ludes, O, Maronas, N. McCallum, D. Moore, N. Morling, H, Niederstatter, F. Noel, W. Parson, C. Popielarz, C. Rapone, A.D. Roeder, Y. Ruiz, E. Sauer, P.M. Schneider, T. Sijen, D. Syndercombe Court, B. Sviezena, M. Turanska, **A. Vidaki**, L. Zatkalikova, J. Ballantyne (2013) RNA/DNA co-analysis from human saliva and semen stains: Results of a third collaborative EDNAP exercise, *Forensic Science International: Genetics* **7(2),** 230-239.

C. Haas, E. Hanson, M.J. Anjos, K.N. Ballantyne, R. Banemann, B. Bhoelai, E. Borges, M. Carvalho, C. Courts, G. De Cock, K. Drobnic, M. Dotsch, R. Fleming, C. Franchi, I. Gomes, G. Hadzic, S.A. Harbison, J. Harteveld, B. Hjort, C. Hollard, P. Hoff-Oslen, C. Huls, C. Keyser, O, Maronas, N. McCallum, D. Moore, N. Morling, H, Niederstatter, F. Noel, W. Parson, C. Phillips, C. Popielarz, A.D. Roeder, L.Salvaderi, E. Sauer, P.M. Schneider, G. Shanthan, D. Syndercombe Court, M. Turanska, R.A.H. van Oorschot, M. Vennemann, **A. Vidaki**, L. Zatkalikova, J. Ballantyne (2014) RNA/DNA co-analysis from human menstrual blood and vaginal secretion stains: Results of a fourth and fifth collaborative EDNAP exercise, *Forensic Science International: Genetics* **8(1),** 203-212.

M. van den Berge, A. Carracedo, I. Gomes, E.A.M. Graham, C. Haas, B. Hjort, P. Hoff-Olsen, O. Maronas, B. Mevag, N. Morling, H. Niederstatter, W. Parson, P.M. Schneider, D. Syndercombe Court, **A. Vidaki**, T. Sijen (2014) A collaborative European exercise on mRNA-based body fluid/skin typing and interpretation of DNA and RNA results, *Forensic Science International: Genetics* **10,** 40-48.

C. Haas, E. Hanson, R. Banemann, A.M. Bento, A. Berti, A. Carracedo, C. Courts, G. De Cock, K. Drobnic, R. Fleming, C. Franchi, I. Gomes, G. Hadzic, S.A. Harbison, B. Hjort, C. Hollard, P. Hoff-Olsen, C. Keyser, A. Kondili, O. Maronas, N. McCallum, P. Miniati, N. Morling, H. Niederstätter, F. Noel, W. Parson, M.J. Porto, A.D. Roeder, E. Sauer, P.M. Schneider, G. Shantan, T. Sijen, D. Syndercombe Court, M. Turanska, M. van den Berge, M. Vennemann, **A. Vidaki**, L. Zatkalikova, J. Ballantyne, (2015) RNA/DNA co-analysis from human skin and contact traces: Results of a sixth collaborative EDNAP exercise, *Forensic Science International: Genetics* **16**, 139-147.

Work in progress:

A. Vidaki, D. Ballard, A. Aliferi, T. Miller, L. Barron, D. Syndercombe Court: DNA methylation-based age prediction using artificial neural networks and next generation sequencing

A. Vidaki, C. Johansson, D. Syndercombe Court: Potential discrimination between whole and menstrual blood using the differentially methylated embryonal Fyn-associated substrate (EFS) gene

A. Vidaki, F. Giangasparo, D. Syndercombe Court: Identification of semen-specific differentially methylated CpG sites using bisulphite pyrosequencing

A. Vidaki, F. Giangasparo, D. Syndercombe Court & EuroForGen partners: Potential forensic tissue identification using immune cell-specific differentially methylated CpG sites

Review Article in Peer-Reviewed Scientific Journal

A. Vidaki, B. Daniel, D. Syndercombe Court (2013) Forensic DNA methylation profiling – Potential opportunities and challenges, *Forensic Science International: Genetics* 7(5), 499-507.

Invited Book Chapter

L. Kovatsi, **A. Vidaki**, D. Fragou, D. Syndercombe Court (2015) 'Epigenetic 'fingerprint'' in Personalised Epigenetics, ed T. Tollefsbol, Elsevier.

Invited Presentations at Scientific Conferences

Is age 'written' in your blood?, Ageing Summit 2016, EuroSciCon, February 2016, London, UK

Invited Seminars at Scientific Institutions

DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing, Analytical and Environmental Sciences Division Seminar Series, October 2015, King's College London, UK

Oral Presentations at Scientific Conferences & Meetings

DNA methylation-based age prediction using artificial neural networks and next generation sequencing, 26th World Congress of the International Society of Forensic Genetics, September 2015, Kraków, Poland

mRNA- and DNA methylation-based differentiation between peripheral and menstrual blood, DNA Conference of the Chartered Society of Forensic Sciences, April 2015, Birmingham, UK

Forensic DNA methylation profiling as an innovative investigative tool, Forensic Horizons R&D Conference of the Forensic Science Society, November 2013, Manchester, UK

DNA methylation-based identification of forensically related body fluids, 2nd EuroForGen General Assembly Meeting, January 2013, Tenerife, Spain

DNA-based estimation of biological age using age-associated DNA methylation markers in blood, Advances in Temporal Forensic Investigations Conference, November 2012, Huddersfield, UK

DNA methylation-based identification of forensically related body fluids, Inaugural Postgraduate Research Symposium of the Forensic Science Society, November 2012, Coventry, UK

Poster Presentations at Scientific Conferences & Meetings

Identification of sperm-specific DNA methylation markers using bisulphite pyrosequencing, 25th World Congress of the International Society of Forensic Genetics, September 2013, Melbourne, Australia

Potential age determination using age-associated DNA methylation markers in blood, 25th World Congress of the International Society of Forensic Genetics, September 2013, Melbourne, Australia

Identification of tissue-specific DNA methylation markers for use in forensic body fluid identification, Graduate School day of Blizard Institute, Barts and the London, April 2012, London, UK